Modeling Temporal Dynamics in Functional Brain Connectivity

Nielsen, Søren Føns Vind

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Modeling Temporal Dynamics in Functional Brain Connectivity

Søren Føns Vind Nielsen

Kongens Lyngby 2018
PHD-2018-484

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Richard Petersens Plads, building 324,
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk
PHD-2018-484

Summary (English)

This thesis deals with modeling temporal changes in functional brain connectivity derived from functional magnetic resonance imaging (fMRI). These changes, observed in both task and rest settings, have been coined *dynamic functional connectivity* (dFC), and are often clustered into a discrete set of so-called dFC *states*. In the five included research papers, we analyse these repeating patterns of connectivity using Bayesian machine learning methods and relate these to cognitive traits and disease status in different resting-state datasets.

In dFC state models, we are faced with many parameter choices, which we in this thesis have tackled using a predictive likelihood framework allowing for quantitative model comparison. Furthermore, this can also be used to assess the relative plausibility of a set of candidate models. We applied this framework to the Wishart mixture model, a probabilistic extension of the sliding-window k-means approach used in many dFC studies. Here, we show that the predictive likelihood can be used to quantify the support for dFC given different window lengths. Furthermore, in another paper we show that the predictive likelihood can be used to choose both the number of states and the model structure in a hidden Markov model (HMM) applied to a highly sampled single subject's resting-state fMRI data.

Another way to investigate the relevance of dFC models is to relate them to subject specific cognitive traits or disease status. The former was investigated in a large cohort of healthy subjects' resting-state fMRI data and we found almost no association between the temporal characteristics of the dFC models and the higher order cognitive traits. In another paper we investigated different HMMs ability to distinguish between patients with schizophrenia and healthy controls

based on resting-state fMRI data. We found that the simplest characterizations using static FC were adequate for the classification task.

Our findings underline the importance of quantitative evaluation of dFC models and furthermore shows that we need better models that can account for subject variability and noise confounds.

Summary (Danish)

Denne afhandling beskæftiger sig med modellering af tidslige ændringer i funktionelle hjerneforbindelser målt ved funktionel magnetisk resonans scanningsbilleder (fMRI). Disse ændringer, der observeres både i den aktive og hvilende hjerne, er blevet navngivet dynamisk funktionel konnektivitet (dFC) og bliver ofte grupperet i et diskret antal af såkaldte dFC *stadier*. I fem forskningsartikler, der danner grundlaget for denne afhandling, analyserer vi disse gentagne mønstre af hjernekonnektivitet ved hjælp af Bayesianske maskinlæringsmetoder og relaterer disse til kognitive træk og sygdomsstatus i forskellige populationers hviletilstands fMRI.

I modeller af dFC-stadier står vi ofte overfor mange parametervalg, som vi i denne afhandling har håndteret ved hjælp af den prædiktive likelihood, der muliggør kvantitativ model sammenligning. Desuden kan dette også bruges til at vurdere den relative plausibilitet af forskellige kandidatmodeller. Denne metode er blevet anvendt på Wishart mikstur modellen, en probabilistisk udvidelse af en vinduesmodel anvendt i mange dFC undersøgelser. Her viser vi, at den prædiktive likelihood kan bruges til at kvantificere graden af dynamik givet forskellige vindueslængder. Desuden viser vi i en anden artikel, at den prædiktive likelihood kan bruges til at vælge både antallet af stadier og modelstrukturen i en skjult Markov model (HMM), der er anvendt på data fra én hvilende persons gentagne fMRI-målinger.

En anden måde at undersøge relevansen af dFC-modeller på er at relatere dem til personspecifikke kognitive træk eller sygdomsstatus. Det førstnævnte blev undersøgt i en stor gruppe af sunde individers hviletilstands fMRI-data, og vi fandt næsten ingen sammenhæng mellem de tidsmæssige egenskaber ved dFC-

modellerne og de højere kognitive træk. I en anden artikel undersøgte vi forskellige HMMs evne til at skelne mellem patienter med skizofreni og raske kontrolpersoner baseret på hvilestilstands fMRI data. Vi fandt ud af, at de simpleste modelstrukturer baseret på statisk FC var tilstrækkelige til klassifikationsopgaven.

Vores resultater understreger betydningen af kvantitativ evaluering af dFC-modeller og viser desuden, at vi har brug for bedre modeller, som kan tage højde for variabilitet mellem personer og anvende mere realistiske støjmodeller.

Preface

This thesis was prepared at the Section for Cognitive Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark in fulfillment of the requirements for acquiring the Ph.D. degree in engineering. The project was funded partly (54%) by the Lundbeck foundation through the Non-parametric Relational Modeling of Functional and Structural Brain Connectivity project (see also https://brainconnectivity.compute.dtu.dk/) and partly by the DTU Compute Phd School. Throughout the project my main supervisor was associate professor Morten Mørup, DTU Compute, Technical University of Denmark. Furthermore, I was co-supverised by associate professor Mikkel N. Schmidt from DTU Compute and associate professor Kristoffer H. Madsen from Danish Research Centre for Magnetic Resonance, Copenhagen University Hospital.

The thesis is paper-based containing five research publications, one journal paper and four peer-reviewed conference publications. The work was carried out between August 2015 and August 2018.

Lyngby, 14-August-2018

Søren Føns Vind Nielsen

List of publications

List of papers **included** in the thesis

Nielsen, Søren F V, Kristoffer H Madsen, Rasmus Røge, Mikkel N Schmidt, and Morten Mørup (2016). "Nonparametric Modeling of Dynamic Functional Connectivity in fMRI Data". In: *NIPS 2015 Workshop on Machine Learning and Interpretation in Neuroimaging*. Montreal, Canada: arxiv.org.

Nielsen, Søren F V, Kristoffer H Madsen, Mikkel N Schmidt, and Morten Mørup (2017). "Modeling dynamic functional connectivity using a wishart mixture model". In: *2017 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*. Toronto, Canada: IEEE, pp. 1–4. DOI: 10.1109/PRNI.2017.7981505.

Nielsen, Søren F V, Mikkel N Schmidt, Kristoffer H Madsen, and Morten Mørup (2018). "Predictive assessment of models for dynamic functional connectivity". en. In: *Neuroimage* 171, pp. 116–134. ISSN: 1053-8119, 1095-9572. DOI: 10.1016/j.neuroimage.2017.12.084.

Nielsen, Søren F V, Yuri Levin-Schwartz, Diego Vidaurre, Tulay Adali, Vince D Calhoun, Kristoffer H Madsen, Lars Kai Hansen, and Morten Mørup (2018). "Evaluating models of dynamic functional connectivity using predictive classification accuracy". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Calgary, AB, Canada.

Nielsen, Søren F V, Diego Vidaurre, Kristoffer H Madsen, Mikkel N Schmidt, and Morten Mørup (2018). "Testing group differences in state transition structure of dynamic functional connectivity models". In: *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*. Singapore.

List of papers **not included** in the thesis

Hinrich, Jesper L, Søren F V Nielsen, Kristoffer H Madsen, and Morten Mørup (2016). "Variational group-PCA for intrinsic dimensionality determination in fMRI data". In: *2016 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*. ieeexplore.ieee.org, pp. 1–4. DOI: 10.1109/PRNI.2016.7552357.

Hinrich, Jesper L, Søren F V Nielsen, Nicolai A B Riis, Casper T Eriksen, Jacob Frøsig, Marco D F Kristensen, Mikkel N Schmidt, Kristoffer H Madsen, and Morten Mørup (2017). "Scalable Group Level Probabilistic Sparse Factor Analysis". In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, LA, USA.

Stevner, Angus B A, Diego Vidaurre, Joana Cabral, Kristina Rapuano, Søren F V Nielsen, Enzo Tagliazucchi, Helmut Laufs, Peter Vuust, Gustavo Deco, Mark W Woolrich, Eus Van Someren, and Morten L Kringelbach (2018). "Discovery of key whole-brain transitions and dynamics underlying the human non-REM sleep cycle". In: Under review.

Jørgensen, Philip J H, Søren F V Nielsen, Jesper L Hinrich, Mikkel N Schmidt, Kristoffer H Madsen, and Morten Mørup (2018). "Probabilistic PARAFAC2". In: arXiv: 1806.08195 [stat.ML].

Acknowledgements

First of all, I would like to thank my supervisors, Kristoffer, Mikkel and Morten, for their guidance, fantastic discussions about science (and other stuff) and continued support even at times when I was being unreasonable. Furthermore, I would like to thank everyone at the Cognitive Systems section at DTU Compute, where most of my time has been spent. Thank you for all the laughs and coffee breaks!

I would also like to thank the Lundbeck foundation for their funding which made this Ph.d. possible, along with the Otto Mønsted and Oticon foundations for their monetary support for my travels to conferences and stay abroad at Oxford University. From my time in Oxford, I would like to thank all the members of the Hedonia lab and OHBA for welcoming me with open arms and for showing me a fantastic time.

Finally, I would like to thank friends and family for their continued support and patience during times when my mind was occupied by machine learning.

Contents

Chapter 1

Introduction

Understanding the human brain and its functional organization is (still) one of the largest scientific challenges of our time. Determining how the brain's mechanisms modulate and shape cognition can help us understand ourselves, both at an individual level and as a species (population level). Enriching this understanding can furthermore provide a basis for diagnosing and treating different neurological conditions, such as schizophrenia, post-traumatic stress, bipolar disease, to mention a few. Since its invention in the early 1990's, functional magnetic resonance imaging (fMRI) (Ogawa et al., 1992; Kwong et al., 1992) has become a popular research tool for studying cognition. fMRI measures (non-invasively) the blood oxygen level dependent (BOLD) signal across the whole brain with a spatial resolution typically in the range of 1-4mm^3. The BOLD signal represents the local oxygen demand in each voxel and is thus only an indirect measure of neuronal firing (related through the haemodynamic response function). Compared to other non-invasive modalities such as electroencephalography (EEG) and magnetoencephalography (MEG), fMRI has a relatively poor temporal resolution, which is due to the slow varying nature of the BOLD-response to neuronal firing.

Most fMRI studies that have followed after its invention are based on the subtraction principle. In its simplest form, a subject is asked to perform task A and task B while inside a scanner (e.g. hand movement and rest) after which a *contrast* is applied that tells what voxels displayed larger BOLD values in task

A relative to task B. This has been a quite successful strategy to map certain cognitive concepts to a location in the brain, but also has the problematic assumption of the tasks (A and B) not interacting in a systematic way (Friston, Price, et al., 1996). Surprisingly, it has been observed that even in a task-free paradigm, denoted the *resting state*, that the BOLD signal is structured by *functional connectivity* between regions (Biswal et al., 1995; M. D. Fox and Raichle, 2007; Smith, P. T. Fox, et al., 2009), i.e. a pattern of co-variation in time. The underlying reason is still not very well understood and being able to characterize this structure can lead to a better understanding of the brain's functional organization.

1.1 Functional Brain Connectivity and the Resting-state

One of the first studies that investigated the role of functional brain connectivity in the resting-state was carried out by Biswal et al., 1995. In their seminal work, they analysed resting-state fMRI data from 11 healthy subjects, that also underwent a functional localizer scan based on self-paced finger-tapping. It was found that the voxels identified from the localizer scan (mostly in motor cortex and supplementary motor area) had a high within-region temporal correlation in the resting-state compared to voxels from outside those regions. This suggests that a functional connection exist between areas with the same task-activation even in a resting-state condition. Furthermore, Smith, P. T. Fox, et al., 2009 showed, using a meta-analysis approach with independent component analysis (ICA) on task-contrast maps and resting-state fMRI data, that the brain's functional organization during a variety of tasks is preserved in the resting-state.

In addition to the observation of functional connections at rest, a task-negative network, coined the *default mode network* (DMN), was discovered by Raichle et al., 2001. They found, using positron emission tomography (PET), a set of regions that exhibited above mean blood flow in the resting-state. However, the analysis revealed a remarkably uniform distribution of brain activity in the resting-state, as measured by the oxygen extraction fraction (OEF), suggesting that the DMN is not "active" in the classical sense and rather represents a baseline metabolic state. The DMN observed here had large overlap with spatial maps from the PET literature associated with activity decrease in attention demanding tasks. However, the inherently stationary assumption of the brain's functional connectivity in the resting-state has been challenged in recent years.

1.2 Resting-state Dynamic Functional Connectivity

One of the first studies to address the non-stationarity of resting-state FC was carried out by Chang and Glover, 2010. They used a wavelet transform method to analyse the time-frequency coherence changes between the posterior cingulate cortex (PCC) and other regions of interest (ROIs). These ROIs were nodes from the DMN and some regions displaying consistent (group-level) negative temporal correlation with the PCC, denoted the anti-correlated nodes. It was observed that the magnitude of coherence fluctuated over time which showed small time-segments of negative correlation between the ROIs considered. It was concluded that this variation could cause the simple linear correlations approaches to FC analysis to oversimplify the temporal dynamics observed.

Allen et al., 2014 investigated the temporal variability of FC in a resting-state fMRI from a population of 405 healthy young adults. A group independent component analysis (gICA)(Calhoun, Adali, et al., 2001) was applied to the entire population to bring subjects into a common lower-dimensional subspace after which a sliding-window procedure was used to extract FC estimates at each point in time. Subsequently, a k-means clustering was applied to obtain a set of *FC states* displaying common FC patterns that reoccur in time. The entire procedure will in this thesis be denoted the sliding-window k-means (SWKM). In the final set of seven FC states it was found that certain connectivity states deviated from the stationary FC, however, with lower temporal occurrence than states resembling static FC. Furthermore, it was illustrated that the FC states differed mostly in their connectivity within and to the DMN challenging the notion of one stable DMN. For reviews on dynamic functional connectivity see (Hutchison et al., 2013; Calhoun, R. Miller, et al., 2014).

Several studies have since the first observations of FC (Biswal et al., 1995) gone beyond the correlational approach and modeled the causal relationships between neuronal systems, denoted *effective connectivity* (Friston, 2011). The most widely used approach in this field is the dynamic causal model (DCM) (Friston, Harrison, et al., 2003), that given a task modulation characterizes the effective connections and their changes between brain areas. In the standard setting, DCM models are used to model dynamical changing effective connectivity induced by behavioral paradigms. However, in DCM the paradigm sequence is predefined and the connectivity is static within each behavioral state. Effective connectivity and models thereof will not be considered in this thesis.

Explanations and hypotheses for fluctuating states of FC in the resting brain come in colors of many. In healthy populations, some of the variability ob-

served in resting-state dFC has been attributed to mind-wandering (Kucyi, 2017); the process in which the mind wanders freely undirected by cognitive control, also sometimes referred to as "daydreaming". Furthermore, in clinical populations, dFC differences (as compared to healthy controls) have been observed in schizophrenia (Ma et al., 2014; Du et al., 2017) and post-traumatic stress disorder (Ou et al., 2015).

However, critique has arisen in the community both in terms of the statistical methodology and also whether the origins of the observed dynamics are in fact neural. On the methodological side, certain parameter choices, such as the window length in SWKM, have been shown to greatly impact the results (Leonardi and Van De Ville, 2015; Zalesky and Breakspear, 2015; Shakil et al., 2016). A reasonable range for the window length has been described by Leonardi and Van De Ville, 2015; Zalesky and Breakspear, 2015, based on the expected frequency content of the BOLD signal. However, there is still need to refine this range as the selection of the window length poses a trade-off between expressiveness (short windows) vs. stability (long windows). In addition to this, the reliability of resting-state dFC has been investigated by Choe et al., 2017, which was measured in a test-retest scenario in resting-state data from the Human Connectome Project (HCP) (Smith, Beckmann, et al., 2013). Summary statistics, such as the mean and variance of the FC time-courses, were compared across the two datasets, which showed a low reliability. Furthermore, the dynamic nature of resting-state FC has been questioned in (Laumann et al., 2016). Here, a synthetic study was carried out in which data with stationary FC was simulated but matched in power spectrum to real resting-state fMRI data. It was observed that the standard SWKM still found highly varying states of connectivity in accordance with those found on real data. Laumann et al., 2016 suggests using higher order moments, *multivariate kurtosis* in this case, to asses the non-stationarity of resting-state FC. Multivariate kurtosis was shown to increase when the simulated data included segments of task (i.e. true cognitive modulation), however, the increase in higher order statistics was also linked to head motion and drowsiness. Overall, most methods for analysing dFC are not rooted in a probabilistic framework that can quantify parameter uncertainty. This makes objective comparison of different models, including choosing certain model parameters, very difficult, further hampering the interpretation of dFC.

1.3 Outline

This thesis aims to model temporal fluctuations in functional connectivity in a data-driven manner. We approach this through a generative model and use the Bayesian inference scheme as a principled means to estimate parameters taking

into account prior knowledgde about the data-generating process. We will first embed the sliding window k-means (SWKM) in probabilistic modeling using a natural Bayesian prior over covariance matrices, leading to the Wishart mixture model (WMM). Furthermore, we will be using the hidden Markov model (HMM) as a alternative window-free approach, that is more general, as it through the choice of emission model can characterize different modeling assumptions on dFC. In preliminary work (Nielsen, 2015), we investigated non-parametric HMM modeling of dFC states within a principal component analysis (PCA) representation of single-subject data demonstrating ability to discriminate between task and rest. However, what drove this difference was unclear. The use of Bayesian modeling lets us tap into features such as held-out predictions, which will be used to quantify the goodness-of-fit of different model structures and model complexity. Next, we will explore if the dynamic structure can characterize disease-state induced differences in subjects (i.e. patients with schizophrenia vs. healthy controls) better than static models. Finally, we will also quantify if dynamic transitions relate to behavioral traits in healthy populations.

The overarching theme of the thesis will be to what degree the fMRI data currently collected supports dFC. This will be addressed considering the following research questions,

- How can we determine the support for different dFC models in a quantitative way?
  - Investigated in (Nielsen, Madsen, Schmidt, et al., 2017; Nielsen, Schmidt, et al., 2018)
- What modeling choices, i.e. the number of states and the parameterization thereof, impact the conclusions drawn from dFC analysis?
  - Investigated in (Nielsen, Schmidt, et al., 2018)
- What factors, i.e. subjects, preprocessing and task, influence the state sequence obtained from dFC models?
  - Investigated in (Nielsen, Madsen, Røge, et al., 2016)
- Is there a relation between features of resting-state dFC and subject specific cognitive traits?
  - Investigated in (Nielsen, Levin-Schwartz, et al., 2018; Nielsen, Vidaurre, et al., 2018)

The rest of the thesis is structured as follows. **Chapter 2** includes a short introduction to Bayesian modeling and inference, including a summary of Bayesian dFC models. **Chapter 3** summarizes the five research papers from this thesis. **Chapter 4** presents a short discussion of the findings and a conclusion. All papers that form the foundation of this thesis can be found in Appendix A.

Chapter 2

Bayesian Modeling of Dynamic Functional Connectivity

One of the cornerstones of this thesis is the Bayesian modeling framework. In this chapter a gentle overview of the principles behind Bayesian modeling will be presented alongside its relation to the dynamic functional connectivity literature.

2.1 Bayesian Modeling

The Bayesian modeling philosophy revolves around quantifying uncertainty in parameters of the model by accumulating evidence from observed data. The model uncertainties can be quantified using Bayes rule for conditional probabilities, which when incorporated into the modeling framework has the form,

$$p(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{X})}, \tag{2.1}$$

in which $\boldsymbol{\theta}$ are the parameters of the model, $\boldsymbol{X}$ the observed data, $p(\boldsymbol{\theta}|\boldsymbol{X})$ is denoted the *posterior*, $p(\boldsymbol{X}|\boldsymbol{\theta})$ is called the *likelihood*, $p(\boldsymbol{\theta})$ is the *prior* and $p(\boldsymbol{X})$ is the *evidence*.

Gelman et al., 2014 describes the ideal Bayesian modeling framework in the following three steps, which can be reiterated,

1. Setting up the full model including prior beliefs,
2. Model inference on given data,
3. Evaluation (including a sensitivity analysis of the assumptions of step 1).

The quantity of interest in Bayesian modeling is the posterior, that specifies the distribution over the parameters of the model given the observed data. The model is specified in the prior distribution along with a likelihood function that describes how likely the data $\boldsymbol{X}$ is given a certain set of parameters $\boldsymbol{\theta}$. Note here that both the prior and the posterior are distributions over the parameters, where the prior quantifies prior beliefs and the posterior the updated beliefs after seeing the data. To yield a probability distribution over $\boldsymbol{\theta}$ the numerator in (2.1) is normalized by the evidence, which can be calculated as,

$$p(\boldsymbol{X}) = \int p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{2.2}$$

However, this integral over all possible parameter values is in most practical cases analytically intractable and one must resort to approximate Bayesian methods.

2.1.1 Approximate Inference

Most approximation methods for the posterior distribution of a Bayesian model fall within one of the following two categories; Markov chain Monte Carlo sampling (MCMC) or variational Bayes (VB).

In MCMC the objective is to obtain a representative set of samples from the posterior distribution, such that a multi-dimensional intractable integral of interest can be approximated. This could for instance be the predictive likelihood on held out data given the training data $p(\boldsymbol{X}^*|\boldsymbol{X})$, which then using MCMC can be estimated by,

$$p(\boldsymbol{X}^*|\boldsymbol{X}) = \int p(\boldsymbol{X}^*|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{X})d\boldsymbol{\theta} \approx \frac{1}{S}\sum_{s=1}^{S} p(\boldsymbol{X}^*|\boldsymbol{\theta}^{(s)}), \tag{2.3}$$

in which $\boldsymbol{\theta}^{(s)}$ is a sample from the MCMC algorithm of which we in total have S. For a thorough review of MCMC see Neal, 1993. This is equivalent to averaging

the predictions over each sample from the chain. Now the question becomes how such a set of samples can be obtained. The idea behind MCMC is to create a Markov chain of samples that (after convergence of the chain) come from the true posterior. The samples are first order Markovian, i.e. a new sample in the chain is conditioned on the previous sample. In most cases, it can be proved that in the limit $S \to \infty$ the samples come from the posterior (by proving the so-called *detailed balance* condition), however, in practice it is not known when the chain has converged. A standard way to address this issue is to use burn-in, where a number of initial samples are discarded. Several more advanced tools for MCMC algorithms have been created to help diagnose when the algorithm has converged to the posterior (Gelman et al., 2014). Examples of prominent MCMC approaches are Metropolis-Hastings and Gibbs sampling.

In contrast to MCMC, VB approximates the posterior with a simpler distribution, $Q(\boldsymbol{\theta})$, that is tractable to fit (cf. Blei et al., 2016 for a review on VB methods). Looking at the evidence, taking the natural logarithm and using Jensens inequality yields,

$$\begin{aligned}
\log p(\boldsymbol{X}) &= \log \int p(\boldsymbol{X}, \boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \log \int p(\boldsymbol{X}, \boldsymbol{\theta}) \frac{Q(\boldsymbol{\theta})}{Q(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&\geq \int \log \left(p(\boldsymbol{X}, \boldsymbol{\theta}) \right) Q(\boldsymbol{\theta}) d\boldsymbol{\theta} - \int \log \left(Q(\boldsymbol{\theta}) \right) Q(\boldsymbol{\theta}) d\boldsymbol{\theta}.
\end{aligned} \tag{2.4}$$

The above lower bound is often referred to as the evidence lower bound (ELBO). One can show that maximizing the ELBO is equivalent to minimizing the Kullback-Leibler (KL) divergence between $Q(\boldsymbol{\theta})$ and the true posterior. The KL-divergence can be written and expanded as,

$$\begin{aligned}
D_{KL}(Q(\boldsymbol{\theta}) | p(\boldsymbol{\theta} | \boldsymbol{X})) &= \int \log \left(\frac{Q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} | \boldsymbol{X})} \right) Q(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \int \log \left(\frac{Q(\boldsymbol{\theta}) p(\boldsymbol{X})}{p(\boldsymbol{\theta}, \boldsymbol{X})} \right) Q(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \log p(\boldsymbol{X}) + \int \log(Q(\boldsymbol{\theta})) Q(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&\quad - \int \log(p(\boldsymbol{\theta}, \boldsymbol{X})) Q(\boldsymbol{\theta}) d\boldsymbol{\theta}.
\end{aligned}$$

This shows that the only thing missing from the ELBO in (2.4) is exactly $D_{KL}(Q(\boldsymbol{\theta}) | p(\boldsymbol{\theta} | \boldsymbol{X}))$, thus by maximizing the ELBO we minimize the KL-divergence between our approximation Q and the true posterior. This optimization problem is tractable if the Q-distribution is chosen properly. A popular choice is

a distribution that factorizes over all the parameters, denoted the *mean-field* approximation, in which the update rules are derived from coordinate ascent. VB methods based on the above approach often have simple update rules, however, a downside is that the Q-distribution is often chosen too simple to not fully model the posterior, i.e. the choice of Q-distribution presents itself with a trade-off between modeling the posterior accurately and computational complexity. Furthermore, optimizing the KL-divergence has a tendency to focus on the largest mode of the posterior distribution which can thus underestimate the posterior variance(Bishop, 2006).

As with the approximation in MCMC (2.3), the log predictive likelihood can be approximated in the VB-framework by,

$$\begin{aligned}\log p(\boldsymbol{X}^*|\boldsymbol{X}) &= \log \int p(\boldsymbol{X}^*, \boldsymbol{\theta}|\boldsymbol{X}) d\boldsymbol{\theta} = \log \int p(\boldsymbol{X}^*|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{X}) d\boldsymbol{\theta} \\ &\approx \log \int p(\boldsymbol{X}^*|\boldsymbol{\theta}) Q(\boldsymbol{\theta}) d\boldsymbol{\theta}. \end{aligned} \tag{2.5}$$

There are other strategies to approximating the predictive likelihood in VB (cf. Beal, 2003), however, the above method stays most true to the VB-objective. Furthermore, for some models, both MCMC and VB predictive likelihood is intractable due to certain parameter dependency structures. Such an example is described in the section on the hidden Markov model (section 2.3), where strategies for dealing with this will be discussed.

2.1.2 Model Selection

In most data science applications, it is not known what the true model is (if such model even exists) for the phenomenon we are describing. So naturally, there is a need to compare different candidate models. To exemplify this process, we look at using the Gaussian mixture model (GMM) to model brain states in functional imaging data. Let $\mathbf{x}_t \in \mathbb{R}^V$ be a vector containing the signal from V regions in the brain at time $t = 1 \dots T$. We now assume that each datapoint belongs to one of K states, where the collection of all datapoints assigned to state k has mean $\boldsymbol{\mu}^{(k)}$ and covariance $\boldsymbol{\Sigma}^{(k)}$. This can be modeled using the

GMM, described by the following generative model,

$$\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha}) \tag{2.6}$$
$$z_t \sim \mathrm{Cat}(\boldsymbol{\pi}), \quad t = 1 \dots T \tag{2.7}$$
$$\boldsymbol{\Sigma}^{(k)} \sim \mathcal{W}^{-1}(\boldsymbol{\Sigma}_0, \nu), \quad k = 1 \dots K \tag{2.8}$$
$$\boldsymbol{\mu}^{(k)} \sim \mathcal{N}(\boldsymbol{\mu}_0, \lambda\boldsymbol{\Sigma}^{(k)}), \quad k = 1 \dots K \tag{2.9}$$
$$\mathbf{x}_t \sim \mathcal{N}(\boldsymbol{\mu}^{(z_t)}, \boldsymbol{\Sigma}^{(z_t)}), \quad t = 1 \dots T, \tag{2.10}$$

in which $\boldsymbol{\pi}$ is a K-dimensional probability vector describing the relative size of each state drawn from a Dirichlet distribution ($\mathrm{Dir}(\cdot)$) with concentration vector $\boldsymbol{\alpha}$, $\mathbf{z}$ is the state sequence which is a vector of length T containing the assignment of each datum to one of the K states drawn from a categorical distribution ($\mathrm{Cat}(\cdot)$), $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}^{(k)})$ is the normal prior for the state means[1] and $\mathcal{W}^{-1}(\boldsymbol{\Sigma}_0, \nu)$ is the inverse-Wishart prior on the covariances.

This yields multiple model selection problems, e.g. the number of states used and the value of λ in the prior on $\boldsymbol{\mu}^{(k)}$. The most natural way to answer these questions in the Bayesian framework is to use the model evidence from (2.2). The model evidence expresses the plausibility of data being generated by the model which can be compared across models. In (2.2) the parameters in the model are integrated out, which naturally penalizes more complex models due to the larger model space integrated over. However, as already stated previously this integral is often intractable. And even if it is tractable we still have the problem of model mismatch, i.e. the assumptions of the model are violated in the data, which can make the model evidence misleading (Bishop, 2006). An example of model mismatch in the context of the GMM could be that the noise in the data had temporal correlation, which is currently not modeled in the above formulation (more on this in the bottom of this section). A popular approximation to the evidence is the Bayesian information criterion (BIC) (Schwarz, 1978), which is derived using the Laplace approximation. The BIC consists of two terms, an in-training-sample log-likelihood term and a complexity term, punishing methods with many parameters. Other model selection criteria exist, such as the Akaike information criterion (AIC) (Akaike, 1974) or the widely applicable information criteria (WAIC) (Watanabe, 2010), that have a similar structure as the BIC. The criteria mentioned here have the upside that they are computationally cheap to calculate, as compared to cross-validation (see below). Each measure has it's own assumptions and asymptotic behaviours, however, all of them have the same model mismatch issues as stated above.

Another approach to the model selection problem is to use cross-validation. Here, it is directly estimated how a certain model choice would generalize to an

[1] The choice of the prior on $\boldsymbol{\mu}^{(k)}$ is done for computational convenience as it allows for analytical marginalization of the mean and covariance of each state.

independent dataset. In cross-validation, a training set and a test set is sampled after which the model is trained on the training set and finally evaluated on the test set. Now, this process is repeated a number of times, where each training-test split is denoted a fold, such that each datapoint in the entire set has been in the test set exactly once. The average generalization (over folds) is then used as a model selection criteria. A major drawback of cross-validation is the computation time involved as this increases by a factor equal to the number of folds. Furthermore, when working with time-series data, having an independent training and test set becomes more tricky as compared to non-temporal data (Bergmeir et al., 2018). For an unsupervised model, like the GMM, the most natural generalization measure is the expected utility(Good, 1952; Piironen and Vehtari, 2017), which can be defined as,

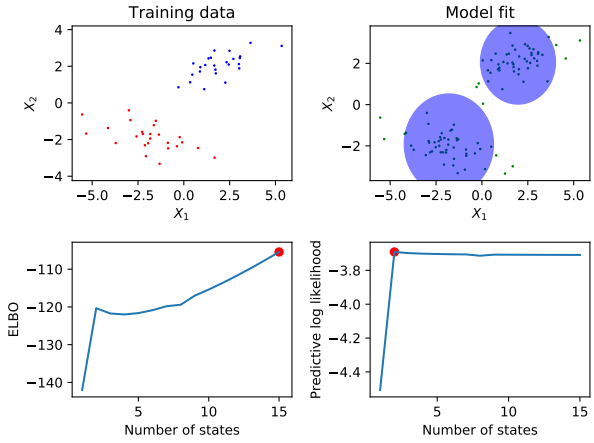
$$\mathbb{E}[\log p(\boldsymbol{X}^*|\boldsymbol{X}, \mathcal{M})] = \int p(\boldsymbol{X}^*) \log p(\boldsymbol{X}^*|\boldsymbol{X}, \mathcal{M}) d\boldsymbol{X}^*, \tag{2.11}$$

in which $\boldsymbol{X}$ is the training set, $\boldsymbol{X}^*$ is a future observation and $\mathcal{M}$ is the candidate model used. Here we have used the most widely applied utility, namely the predictive log-likelihood, as maximizing (2.11) also minimizes the KL-divergence between the true data generating process, $p(\boldsymbol{X}^*)$, and the predictive distribution. In (2.11), all future observations are integrated out, which (obviously) is not manageable in pratice. However, this can be approximated in a cross-validation setting (Geisser and Eddy, 1979) by,

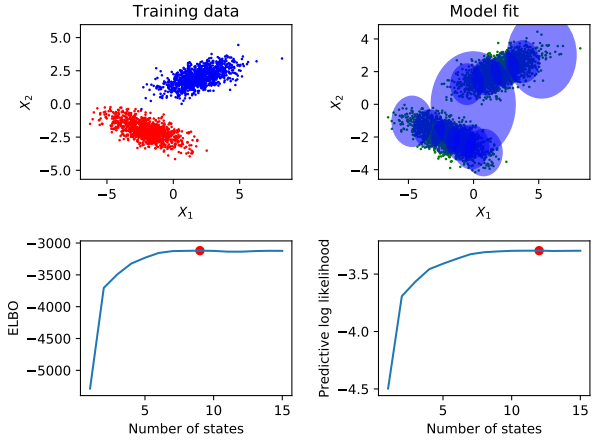
$$\mathbb{E}[\log p(\boldsymbol{X}^*|\boldsymbol{X}, \mathcal{M})] \approx \frac{1}{N} \sum_{n=1}^{N} \log p(\boldsymbol{X}^*_{I(n)}|\boldsymbol{X}_{I(n)}, \mathcal{M}), \tag{2.12}$$

in which $\boldsymbol{X}_{I(n)}$ and $\boldsymbol{X}^*_{I(n)}$ denote the training and test set in the n'th cross-validation split of which there is N.

We illustrate the challenges of the model selection problem with the GMM using the VB-GMM (`BayesianGaussianMixture`) implementation in Scikit-learn (Pedregosa et al., 2011). We generate data from a GMM with two clusters, each having a mean and a full covariance matrix. Now, we introduce a model mismatch by trying to fit a GMM with spherical covariance and at the same time estimate the number of clusters. The results can be seen in Figure 2.1 for few ($T = 50$) and many datapoints ($T = 2000$) respectively. In the case of few samples, we see that the (training) ELBO points to a lot of states due to the fact that it can fit the training data near perfectly. The predictive likelihood on the other hand is more conservative since the extra clusters do not generalize well. When we have many samples in the data (in this case $T = 2000$ in Figure 2.1b), we see that the two model selection criteria are more aligned. However, as expected, both of them do not point toward the "true" number of underlying states due to the clear model mismatch.



(a) Few number of samples for training ($T = 50$)



(b) Large number of samples for training ($T = 2000$)

Figure 2.1: Assessment of model complexity under model mismatch. We generated data ($T = 50, 2000$) from a two-state GMM with mean and full covariance. A spherical GMM was then fitted for $K = 1{:}15$. Each GMM was restarted ten times and we report the ELBO on the training data and the predictive log-likelihood on held-out test data averaged over the restarts. For fitting the final model on the concatenated training and test data (top right panel) we used the predictive log-likelihood as a model selection criterion.

2.2 Wishart Mixture Model

The sliding-window k-means (SWKM) (Sakoğlu et al., 2010; Allen et al., 2014) approach to estimating dFC states is the most widely used method in the literature. It works by applying a windowing procedure to the data such that a set of V-by-V correlation matrices, $\mathcal{C} = \{\boldsymbol{C}_1, \boldsymbol{C}_2, \ldots, \boldsymbol{C}_L\}$ is obtained, after which a clustering is performed on that set to extract a (preferably small) number of FC states. To embed the clustering step in a probabilistic framework a distribution over correlation matrices is needed; such distributions exist (Pourahmadi and Wang, 2015), however, they are computationally impractical. The Wishart distribution arises naturally as the distribution over the Gramian matrix, $\mathbf{G} = \sum_t \mathbf{x}_t \mathbf{x}_t^T$, of a zero-mean multivariate Gaussian signal, and can thus act as a likelihood in the clustering of FC connectivity matrices with Gram-matrices instead of correlation. The generative model for the Wishart mixture model (Hidot and Saint-Jean, 2010; Korzen et al., 2014; Cherian et al., 2016; Nielsen, Madsen, Schmidt, et al., 2017) can be written as,

$$\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha}) \tag{2.13}$$

$$z_\ell \sim \mathrm{Cat}(\boldsymbol{\pi}), \quad \ell = 1 \ldots L \tag{2.14}$$

$$\eta \sim \mathcal{G}^{-1}(a_0, b_0) \tag{2.15}$$

$$\boldsymbol{\Sigma}^{(k)^{-1}} \sim \mathcal{W}(\eta \mathbf{I}_V, \nu_0), \quad k = 1 \ldots K \tag{2.16}$$

$$\mathbf{C}_\ell \sim \mathcal{W}(\boldsymbol{\Sigma}^{(z_\ell)}, \nu_\ell), \quad \ell = 1 \ldots L, \tag{2.17}$$

in which $\boldsymbol{\pi}$ is a K-dimensional probability vector describing the relative size of each state drawn from a Dirichlet distribution ($\mathrm{Dir}(\cdot)$) with concentration vector $\boldsymbol{\alpha}$, z_ℓ is the state assignment of each FC matrix to one of the K states drawn from a categorical distribution $\mathrm{Cat}(\boldsymbol{\pi})$, η is the strength of the regularization on the FC states drawn from an inverse Gamma distribution with shape and scale a_0 and b_0, $\boldsymbol{\Sigma}^{(k)^{-1}}$ is the k'th FC state's inverse covariance which has a Wishart prior with matrix argument $\eta \mathbf{I}_V$ (scaled identity matrix) and degrees of freedom ν_0. The degrees of freedom in the likelihood, ν_ℓ, is always set to the number of samples used to estimate the observed Gramian matrix (i.e. dependent on the window length).

Inference in this model can be done using expectation-maximization, as originally proposed by Hidot and Saint-Jean, 2010. In the case of Korzen et al., 2014; Cherian et al., 2016 a more advanced prior over the state-assignments was used that allows for an infinite number of states, and in that case MCMC was applied as the inference engine. In Nielsen, Madsen, Schmidt, et al., 2017, we have applied a coordinate ascent variational Bayesian approach that is more computationally efficient than the MCMC-approaches, however, at the cost of some simplifying assumptions, i.e. that the posterior factorizes. The de-

tails of the inference including update rules for all parameters can be found in Nielsen, Madsen, Schmidt, et al., 2017 and the MATLAB code is available at https://github.com/sfvnDTU/vbwmm.

2.3 Hidden Markov Model

A downside of both the Wishart mixture model (WMM) and Gaussian mixture model (GMM) is that they do not fully take into account the temporal dependencies of the data samples and the apriori specification of the window length. The hidden Markov model (HMM) can be seen as an extension of the GMM that assumes that the state assignments are first order Markovian, i.e. that assignment of timepoint t, z_t, is dependent on the preceeding state-assignment, z_{t-1}. This has the advantage that the HMM can model temporal dependencies through the *transition matrix*, $\boldsymbol{\pi}$, quantifying the probability of transitioning from one state to another. However, the inference of the model becomes more tricky and more computationally demanding as compared to a standard mixture model.

Using the same notation as in section 2.1.2, when describing the GMM, we can write the generative model for the Bayesian Gaussian HMM with K states as follows,

$$\boldsymbol{\pi}_0 \sim \text{Dir}(\boldsymbol{\alpha}) \tag{2.18}$$

$$\boldsymbol{\pi}^{(k)} \sim \text{Dir}(\boldsymbol{\alpha}^{(k)}), \quad k = 1 \dots K \tag{2.19}$$

$$z_1 \sim \text{Cat}(\boldsymbol{\pi}_0), \tag{2.20}$$

$$z_t | z_{t-1} \sim \text{Cat}(\boldsymbol{\pi}^{(k)}), \quad t = 2 \dots T \tag{2.21}$$

$$\boldsymbol{\Sigma}^{(k)} \sim \mathcal{W}^{-1}(\boldsymbol{\Sigma}_0, \nu), \quad k = 1 \dots K \tag{2.22}$$

$$\boldsymbol{\mu}^{(k)} \sim \mathcal{N}(\boldsymbol{\mu}_0, \lambda \boldsymbol{\Sigma}^{(k)}), \quad k = 1 \dots K \tag{2.23}$$

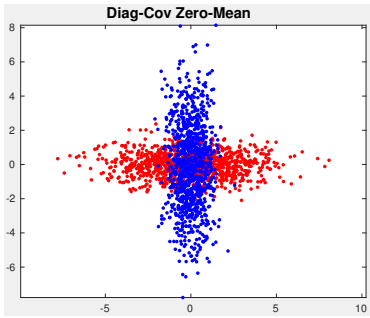
$$\mathbf{x}_t \sim \mathcal{N}(\boldsymbol{\mu}^{(z_t)}, \boldsymbol{\Sigma}^{(z_t)}), \quad t = 1 \dots T, \tag{2.24}$$

in which $\boldsymbol{\pi}_0$ is the K-by-1 vector of initial state probabilities, $\boldsymbol{\pi}^{(k)}$ is the k'th row of the K-by-K transition matrix, z_1 is the state assignment of the first time point, $z_t | z_{t-1}$ denotes the state assignment of timepoint t given the state assignment of the previous timepoint, $\boldsymbol{\mu}^{(k)} \in \mathbb{R}^V$ and $\boldsymbol{\Sigma}^{(k)} \in \mathbb{R}^{V \times V}$ are the parameters governing the distribution of all datums assigned to state k and $\mathbf{x}_t \in \mathbb{R}^V$ is the observed datapoint at time t. Notice if multiple subjects (or sessions) are analysed, discontinuities can arise in the data. This is handled by "restarting“ the state sequence for each subject such that the initial state probability vector is used to draw the state assignment of the first point in the new subject.

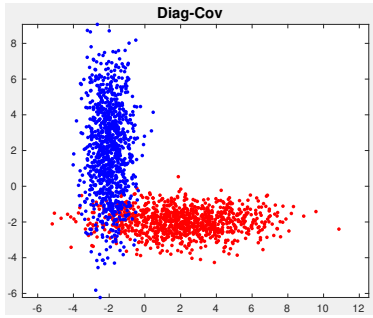
As eluded to in the section on model selection (section 2.1.2), in the GMM there are some parameter choices crucial to the modeling process and this is also true for the HMM. One of them is how the mean and covariance of each state should be parametrized, denoted the *emission model*. Different constraints can be incorporated either due to a specific hypothesis about the data or to bring down the number of estimated parameters in limited sample sizes. In Figure 2.2 six different emission models are visualized for a two-state model that all follow the above generative model with the following modifications; shared covariance over all states (keyword: *stationary*), diagonal covariance (keyword: *diag-cov*) or forcing the mean to be zero (keyword: *zero-mean*) and some combinations thereof. More advanced emission models have been considered in E. B. Fox, 2009; Ryali et al., 2016; Vidaurre, Quinn, et al., 2016 using an vector autoregressive process on the mean function. Furthermore, a very flexible non-parametric emission distribution was considered in De Castro et al., 2016; Kandasamy et al., 2016.

2.3.1 Approximate Inference

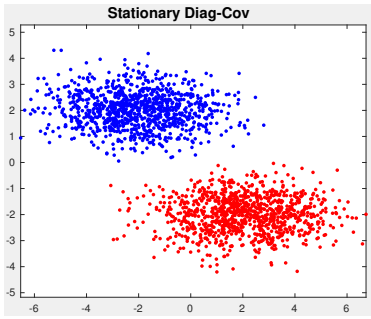
Approximate inference in the model can be carried out both using VB and MCMC. The VB version (Rezek and Roberts, 2005) is very similar to the maximum likelihood algorithm for HMM based on expectation-maximization (EM). In the EM-algorithm for the HMM, an expectation step (E-step) is performed on the state sequence keeping all parameters in the emission model (e.g. mean, $\boldsymbol{\mu}^{(k)}$, and covariance, $\boldsymbol{\Sigma}^{(k)}$ in the likelihood), transition matrix, $\boldsymbol{\pi}$, and initial state probabilities, $\boldsymbol{\pi}_0$, fixed. In the maximization step (M-step) all the fixed parameters from the E-step are updated keeping the state sequence expectation values fixed. When updating the transition matrix and initial state probabilities in the M-step, the expectations $\mathbb{E}[z_t = k]$ and $\mathbb{E}[z_t = j, z_{t-1} = k]$ are needed. An efficient way of calculating these quantities were derived in Baum, 1972, denoted the *Baum-Welch* algorithm, which was later named the *forward-backward* algorithm (Rabiner, 1989). For a detailed derivation of the EM-algorithm for the Gaussian HMM see Bishop, 2006, chapter 13. The VB-version replaces the maximization steps by expectation steps by incorporating a prior (e.g. (2.23) and (2.22)) and doing moment-matching (Bishop, 1999; Rezek and Roberts, 2005). The HMM can be used in a group setting with multiple subjects handling the discontinuities as described above. The forward-backward algorithm can even be run in parallel over subjects making the inference quite fast and efficient. However, in case of analysing large cohorts (like the Human Connectome Project with over 1000 subjects) it becomes quite time consuming to infer the model parameters. A stochastic variational inference HMM for big data applications was developed in Vidaurre, Abeysuriya, et al., 2017, in which a subset of subjects is sampled at random (a *batch*) and the forward-backward algorithm



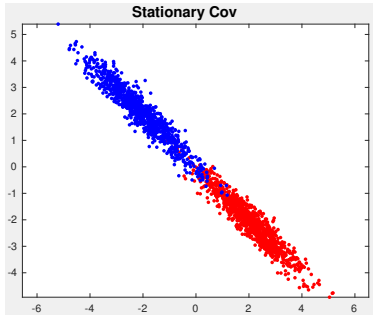
(a) Zero-mean with diagonal covariance



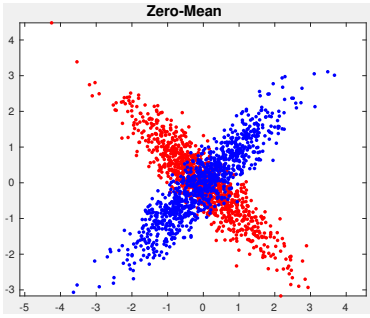
(b) Diagonal covariance



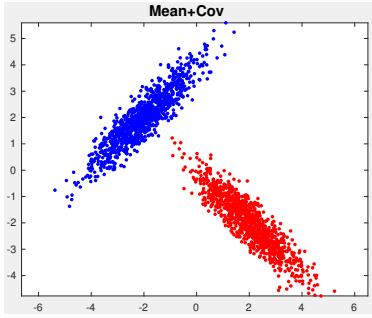
(c) Stationary diagonal covariance



(d) Stationary covariance



(e) Zero-mean



(f) Mean and full covariance

Figure 2.2: Illustration of different emission models (used in Nielsen, Levin-Schwartz, et al., 2018). Six different emission distributions for a two-state model are shown in the above plots.

is run on that subset. A weighted update of the emission model parameters is then performed to reduce the amount of change in parameters due to batch variability (Hoffman et al., 2013). Finally, the subject sampling weights are updated to promote subjects that have not been sampled frequently.

Another approximate inference approach is Markov chain Monte Carlo (MCMC) in which samples are drawn to represent the intractable posterior distribution. The infinite hidden Markov model (iHMM) was proposed by Beal et al., 2002 and in this MCMC sampling was used. The generative model for the iHMM differs from the above given model in that a prior (the Dirichlet Process) is placed on the rows of the transition matrix that allows for a potentially infinite number of states. The inference of the state sequence works by taking one of the datapoints and conditioning on the remaining state assignments. This forms a one-dimensional discrete distribution in which the datum can be assigned to one of the existing states or a new state with some prior probability. This is repeated over the entire state sequence, always maintaining updated summary statistics from the emission model (for more details on this see E. B. Fox, 2009). The vector autoregressive iHMM was implemented in Nielsen, 2015 using a Gibbs sampler with split-merge moves (Jain and Neal, 2004) and temporal noise modeling.

2.3.2 Predictive Likelihood

For model selection in the HMM it can be advantageous to calculate the predictive likelihood on held-out data. The predictive likelihood for the HMM can be written as,

$$p(\boldsymbol{X}^*|\boldsymbol{X}) = \iiiint p(\boldsymbol{X}^*, \mathbf{z}^*, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\pi}_0|\boldsymbol{X}) d\mathbf{z}^* d\boldsymbol{\theta} d\boldsymbol{\pi} d\boldsymbol{\pi}_0, \tag{2.25}$$

in which $\boldsymbol{X}$ is the training set, $\boldsymbol{X}^*$ is the test set, $\mathbf{z}^*$ is the state sequence for the test set, $\boldsymbol{\theta}$ are all relevant parameters for the emission model (for instance $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for each state), $\boldsymbol{\pi}$ is the transition matrix and $\boldsymbol{\pi}_0$ are the initial state probabilities. The above integral includes a marginalization over the state sequence of the test set. However, this is not as straightforward as in a standard mixture model, where the integration over the state assignments is analytically simple. Equation (2.25) can be expanded as,

$$p(\boldsymbol{X}^*|\boldsymbol{X}) = \iiint \sum_{\mathbf{z}^*} p(\boldsymbol{X}^*, \mathbf{z}^*|\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\pi}_0, \boldsymbol{X}) p(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\pi}_0|\boldsymbol{X}) d\boldsymbol{\theta} d\boldsymbol{\pi} d\boldsymbol{\pi}_0, \tag{2.26}$$

replacing the integral over $\mathbf{z}^*$ with a sum. The notation $\sum_{\mathbf{z}^*}$ here denotes the summation over all possible state sequences.

In the variational Bayesian framework (2.26) is approximated by first replacing the posterior with the estimated Q-distribution, and afterwards doing VB-integration of the state sequence (for more details on this see (Beal, 2003) or the appendix of (Nielsen, Schmidt, et al., 2018)). In MCMC a set of S samples are collected from the inference such that (2.26) is approximated by,

$$p(\boldsymbol{X}^*|\boldsymbol{X}) \approx \frac{1}{S}\sum_s \sum_{\mathbf{z}^*} p(\boldsymbol{X}^*, \mathbf{z}^*|\boldsymbol{\theta}^{(s)}, \boldsymbol{\pi}^{(s)}, \boldsymbol{\pi}_0^{(s)}, \boldsymbol{X}), \tag{2.27}$$

in which $\boldsymbol{\theta}^{(s)}, \boldsymbol{\pi}^{(s)}, \boldsymbol{\pi}_0^{(s)}$ denote parameter samples from the MCMC procedure. The summation over all possible state sequences can be handled efficiently using a modified Viterbi-algorithm (cf. technical appendix B.1 or (Nielsen, 2015)).

2.4 Bayesian Dynamic Functional Connectivity in fMRI

Bayesian modeling has, to the best of our knowledge, been used very limited in the context of modeling dynamic functional connectivity (dFC) in fMRI (Ryali et al., 2016; Vidaurre, Smith, et al., 2017; Nielsen, Madsen, Røge, et al., 2016; Nielsen, Schmidt, et al., 2018; Nielsen, Madsen, Schmidt, et al., 2017; Taghia et al., 2018). Here the most recent approaches (not part of this thesis) to modeling dFC using Bayesian hidden Markov models are highlighted, and how each of them deals with the model selection problem.

Ryali et al., 2016 used a variational Bayesian HMM (VB-HMM) (with state individual mean and covariance) to investigate the dFC between three brain systems crucial to human cognition, namely the salience network (SN), central executive network (CEN) and the default mode network (DMN). Their approach was used to investigate brain maturation in young children (age 7-9) as compared to young aduts (age 19-22) (Qin et al., 2012), and afterwards replicated on two adult cohorts from the Human Connectome Project (HCP) (Smith, Beckmann, et al., 2013). The VB-HMM found short lived emerging states of connectivity between SN, CEN and DMN, however, how these states were populated and their lifetimes differed significantly between the adult and child cohort, i.e. children were prone to stay longer in each state and exploring fewer states. The number of states in the VB-HMM was initially set to a high number ($K = 25$), after which a community detection algorithm was applied to the partial correlation matrix from each state. States with the same community structure were then merged to one state, thus reducing the total number of states. In the HCP data, this procedure resulted in 16 and 19 states in the two cohorts, respectively.

Vidaurre, Smith, et al., 2017 used a VB-HMM (with state individual mean and covariance) to analyse whole-brain dFC in the resting-state HCP data. By applying a community detection algorithm to the transition matrix of the estimated HMM they found two sets of states, denoted *metastates*, characterized by higher within metastate transition probability compared to between metastate transition probability. This seperation into metastates was also found by clustering the time spent in each individual state, *fractional occupancy*. Furthermore, it was observed that the difference in fractional occupancy of the two metastates (denoted the metastate profile) for a subject was better predicted by the collection of behavioral traits compared to predicting the fractional occupancy of any of the individual states. This suggests that the metastates are behaviourally relevant features that were also shown to be consistent across sessions within a subject. Regarding the number of states chosen, the analysis was replicated for $K = 8, 12, 16$ states and using split-half, showing that the conclusions about hierarchical state organization are robust.

Taghia et al., 2018 used an extended version of the VB-HMM with an autoregressive process and a factor analysis model for each state in the emission, denoted a Bayesian switching dynamical system (BSDS). The model was used to analyse the working memory (WM) task from the HCP, in which subjects are watching a stream of stimuli and are asked to press a button whenever the current stimuli is the same a the one presented two images before (2-back condition). The task also contains two control conditions; a 0-back condition in which a target stimuli is presented at the beginning of the condition block and a resting-state block in which the subject looks at a fixation cross. The expected state assignment (a 1-by-K vector) from the BSDS model was used a features in an support vector machine (SVM) classifier, to at each time point predict the task condition, i.e. a three way classification problem with classes 2-back, 0-back or rest. The number of states in the BSDS model was set to 8. Over subjects the classification accuracy ranged from $49 - 55\%$, which is better than the chance level of 33%. The subject performance during the 2-back condition, i.e. number of correct responses, was shown to be predicted by the state occupancy of the BSDS states suggesting that engaging in certain brain states results in better WM performance.

Chapter 3

Summary of Research Contributions

Our contributions to the field of dynamic functional connectivity (dFC) revolve around Bayesian modeling of dFC. In Nielsen, Madsen, Røge, et al., 2016, we investigated the use of a non-parametric Bayesian hidden Markov model (HMM), denoted the infinite HMM (IHMM), to model dFC. A hierachical Dirichlet process (Beal et al., 2002; E. B. Fox et al., 2008) was used as a prior on the transition matrix alleviating the need to choose a fixed number of states apriori. This has the upside that the model can adapt to the complexity of data, but comes at a cost that it is sensitive to hyperparameter choices. We investigated the use of the IHMM on fMRI data from 29 healthy subjects from two conditions; a finger-tapping experiment (denoted motor) (Rasmussen et al., 2012) and a resting-state condition (Andersen et al., 2014). We ran an independent component analysis (ICA) using 20 components, discarded six noise components and ran the remaining analysis on the time-courses from these components. We chose a relatively low number of components due to the computational complexity of the IHMM. On a subject-level, we found support for using multiple states to model the two conditions (as opposed to model the data using static FC), and furthermore we could on held out parts of the data predict whether it came from the motor or rest condition with high accuracy. Then we carried out a group-analysis by concatenating data from all the subjects and conditions and ran the IHMM on the entire dataset. We saw using mutual information that the extracted state sequence corresponded more to subject variability and difference

in preprocessing than to task-differences induced by concatenating motor and rest conditions. This indicates that even though we can detect dynamics at a single-subject level that are neurologically meaningful care must be taken when carrying out a group analysis as subject- and preprocessing-variability plays a role.

In Nielsen, Madsen, Schmidt, et al., 2017, we investigated a variational Bayesian formulation of the Wishart mixture model (WMM) (Hidot and Saint-Jean, 2010; Korzen et al., 2014; Cherian et al., 2016) for modeling dFC. Applying a windowing procedure, like the one in Allen et al., 2014, yields a set of correlation matrices. Instead of the widely used k-means algorithm for clustering the correlation matrices into a set of dFC states, we used the WMM that assumes a Wishart likelihood on the matrices. Since the Wishart distribution in statistics naturally arises as the distribution over the Gramian matrix of a zero-mean Gaussian variable, we used Gram matrices instead of correlation. We proposed a heuristic to choose the window length based on the predictive Bayes factor. For each window length, we calculated the predictive log likelihood on held-out data using a range of values for the number of states and contrasted that to the one-state model corresponding to static FC. The window length displaying the largest increase in predictive Bayes factor compared to the one-state model was chosen as the appropriate window length. We saw on synthetic data that we could recover the true window length for different SNR levels, and on an 85 component ICA representation from single-subject resting-state fMRI data (Poldrack et al., 2015), we saw that this heuristic pointed toward a window length of around 30 s. Furthermore, for appropriately chosen window lengths we observed support for multiple states, i.e. the Bayes factor increased when we added more states.

The utility of the predictive likelihood as a model comparison tool was further investigated in Nielsen, Schmidt, et al., 2018. Here, we used the hidden Markov model (HMM) as an example of a framework in which model selection issues exist, due to its recent use in the neuroimaging dFC literature (Vidaurre, Quinn, et al., 2016; Ryali et al., 2016). We investigated three parametrizations of the emission model, each with increasing complexity; a zero-mean Gaussian (ZMG), a Gaussian with state-specific mean (SSM) and a Gaussian with vector-autoregressive mean (VAR). The three models were evaluated using predictive Bayes factors (based on predictive likelihood), in which the baseline model was the static FC. We did this for synthetic data, an electroencephalography (EEG) data set from Wakeman and Henson, 2015 and single-subject fMRI data from Poldrack et al., 2015. From the analysis on synthetic data, we saw that in a case of model-mismatch (as illustrated in section 2.1.2) the simpler parametrizations, not accounting for temporal correlation and smoothness in the data, overestimated the number of states used. To investigate the different methods in a high signal-to-noise ratio setting with high temporal resolution, we fitted the mod-

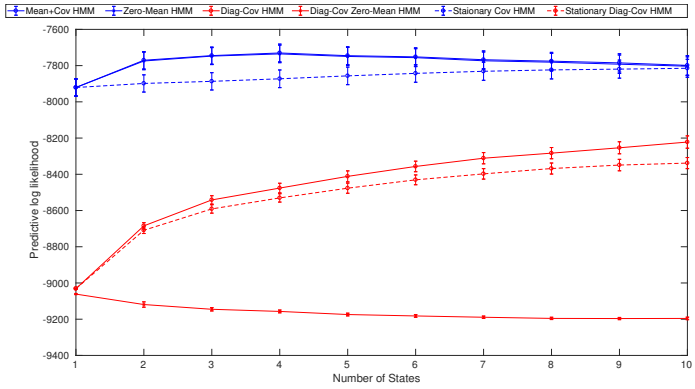
els to an ICA representation (5 components) of event-related potential EEG data from (Wakeman and Henson, 2015). We found that the ZMG and SSM here estimated more states as compared to the VAR, and that the VAR found a consistent baseline state across subjects both pre- and post-stimulus presentation. On single-subject fMRI data using 30 ICA components (discarding 9 components due to spatial overlap with noise sources), we evaluated each of the three HMM parametrizations using predictive Bayes factors and found that the VAR emission model achieved the best performance, notably with support for multiple states. The extracted dynamics were influenced by the different model assumptions, as we observed that the VAR differed in the dynamic state structure as compared to the ZMG and SSM. Furthermore, we made a qualitative comparison of the best HMM solution with a SWKM with the same number of states. Particularly the mean state life-time seemed different across the two methods, where the HMM both had short and long lived states in contrast to the SWKM.

The remainder of our contributions deals with relating dFC models to cognitive traits. In Nielsen, Levin-Schwartz, et al., 2018, we investigated the emission model selection problem in the VB-HMM (Vidaurre, Quinn, et al., 2016), however, this time using predictive classification accuracy. We chose the task of separating patients with schizophrenia (SZ) from healthy controls (HC) based on an 85 component group ICA representation of resting-state fMRI data due to previous results in the literature describing aberrant dynamic functional connectivity in this type of cohort (Ma et al., 2014). The number ICs was reduced to 44 after discarding 37 components due to low fractional amplitude of low-frequency fluctuation (fALFF) (Zou et al., 2008) and four components due to spatial overlap with known noise sources. For each class (SZ and HC), an HMM was trained and the class-membership of a held-out subject was predicted using a Bayes classifier. In training, we varied the emission model and the number of states. The emission models we tested in this paper were chosen to investigate the need for the full covariance when modeling dFC (cf. Figure 2.2 an overview of the emission models used). Our findings suggest that a very simple emission model (few states and diagonal covariance) can capture the differences in FC between SZ and HC. However, this does not mean that we cannot use the more advanced dFC models. To further dive into this[1], we report the predictive log-likelihood for each of the models and each group separately on the the test data in Figure 3.1. Even though the different model parameterizations performed similarly in terms of classification accuracy (cf. Nielsen, Levin-Schwartz, et al., 2018), there is quite a large gap in the predictive log likelihood between the models with full covariance (blue) and the ones with diagonal covariance (red). The above underlines the important distinction between characterization ("what

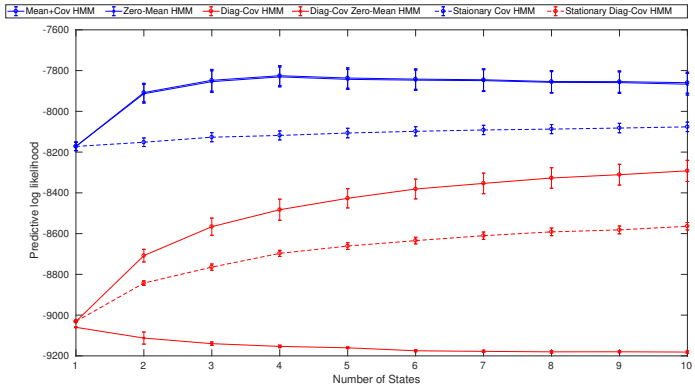
[1]This is a previously unpublished result based on the analysis done in Nielsen, Levin-Schwartz, et al., 2018

is the complexity of the data?") and discrimination ("what features are needed to discriminate two classes apart?").

The relationship between dFC features and behavioural traits in a healthy population was investigated in Nielsen, Vidaurre, et al., 2018. We used the VB-HMM (Vidaurre, Quinn, et al., 2016) and SWKM on resting-state data from the Human Connectome Project (HCP) (Smith, Beckmann, et al., 2013), in which the group ICA representation with 50 compoents was used as in (Vidaurre, Smith, et al., 2017). To stay comparable with previous literature on this matter (Vidaurre, Smith, et al., 2017), we used an HMM with 12 states including a state-specific mean, and likewise a SWKM with 12 states with varying window length. After fitting the models, we divided the subjects into two groups based on different behavioral variables and investigated, using permutation testing, if certain dFC-features were significantly different between the two groups. The dFC-features were fractional occupancy (overall prevalence to a certain state), a measure of state persistency (how likely it is to stay in the same state) and the full transition matrix (the probability to switch between states for all pairs of states). To quantify the difference between two transition matrices obtained from the grouping, we used the total variation measure for probability distributions, as each row of the transition matrix forms a discrete probability distribution. The two "behaviorals" yielding consistent significant differences (over models and dFC features) were gender and a measure of motion in the scanner. This result is fairly unsurprising as gender and motion have previously been linked to dFC differences (Yaesoubi et al., 2016; Laumann et al., 2016). Furthermore, we made the empirical observation that a lot of subjects are needed to find significant differences in the dFC features tested. However, as noted in the paper (Nielsen, Vidaurre, et al., 2018), this can be due to the inappropriate binarization of continuous behavioral variables.



(a) Healthy control group (HC)



(b) Group diagnosed with schizophrenia (SZ)

Figure 3.1: Predictive likelihood analysis of results from Nielsen, Levin-Schwartz, et al., 2018. The predictive log-likelihood (PLL) as a function of the number of states in the HMM model is reported (error-bars indicate standard-error over 10-fold cross-validation). The predictive log-likelihood is calculated separately for each group (HC and SZ), i.e. the HMM model trained on the HC group is only used to calculate the PLL on subjects from the HC group in the test-set. The different model parameterizations correspond to the ones presented in section 2.3.

Chapter 4

Discussion and Conclusion

The advent of machine learning and data science as a core framework in many branches of science has generated a lot of research papers, including the field of neuroimaging. However, many of the current approaches lack a principled way of comparing models thus hampering our ability to establish a state-of-the-art approach. This thesis has described and dealt with probabilistic modeling of dynamic functional brain connectivity (dFC) in functional magnetic resonance imaging (fMRI) data. We have used Bayesian modeling as a central framework to model different dynamical connectivity features of the blood oxygen level dependent (BOLD) signal (obtained from fMRI).

The use of any data-driven method (Bayesian or not) for modeling dFC requires certain parameterization choices, e.g. in the sliding-window k-means (SWKM) approach the window length needs to be specified and the number of brain *states* that will be extracted. Testing different hypotheses about the time-varying nature of FC can be cast as a model selection problem with different candidate models each representing a particular "hypothesis“. In Nielsen, Madsen, Schmidt, et al., 2017, we embedded the SWKM in a Bayesian context using the Wishart mixture model (WMM), and developed a heuristic to choose the window length. This was done by quantifying the generalization as a function of the number of states using predictive log-likelihood for each window length. We saw that both too short and too long windows were not able to utilize additional states added to the model, when we contrasted their

generalization to the static model counterpart. In Nielsen, Schmidt, et al., 2018, we further used the predictive log-likelihood framework to asses the plausibility of different dFC state models. Both our synthetic and real data results show that there is a substantial model dependence in terms of identified dynamic structure, underlining that the dynamic nature of the BOLD-signal (i.e. the number of states) must always be evaluated with the emission model in mind. Especially, the vector-autoregressive (VAR) emission model diverged from the other emission models tested in the state sequence properties. This could be explained by the VAR's ability to capture some of the frequency content in the resting-state BOLD signal (Chang and Glover, 2010; Laumann et al., 2016).

Furthermore, there seems to be a lot of factors that contribute to the FC-variation observed in many studies. Especially in resting-state fMRI, where the large subject variability due to the unconstrained nature of the experiment makes estimation of a common set of neurally relevant dFC states difficult. This was partly observed in (Nielsen, Madsen, Røge, et al., 2016), where we concatenated task and rest to mimick cognitive modulation (i.e. true dynamics) and extracted the state sequence obtained from a group HMM model. The resulting sequence was shown to be driven by subject and session variability to a higher degree than the cognitive modulation from switching between task and rest. Furthermore, we acknowledge that the choice of dimensionality of the subspace, i.e. the number of independent components chosen, has an influence on the generalizability of the estimated dynamics, as stated in Nielsen, 2015 in the context of principal component analysis (PCA). In the research contributions presented in this thesis, we have in some cases chosen the number of components based on computational convenience and in other cases to stay in accordance with previous studies on the same data. The relation between the dimensionality and generalizability in the context of dFC needs to be investigated further.

From a Bayesian modeling point of view, the most elegant way of handling how many states to estimate is by introducing a prior over the state sequence that allows for a potentially infinite number of states. Such an approach has been suggested for the HMM in Beal et al., 2002, denoted the infinite HMM (iHMM). From a cognitive neuroscience perspective, this model also has many desirable properties, including the ability to model new "cognitive" states emerging as we include more data. Especially in resting-state fMRI, where the unconstrained nature of the experiment calls for models that account for this, we see a potential use for non-parametric models that can adapt their complexity. However, in practice we saw during a number of our experiments (Nielsen, Madsen, Røge, et al., 2016; Nielsen, Schmidt, et al., 2018) that the iHMM was sensitive to certain prior-settings (e.g. the regularization on the FC-states), which in turn made it necessary to cross-validate for the optimal settings. Furthermore, the parametric version of the HMM seemed to give similar results in terms of the temporal properties as the iHMM after cross-validation.

The model selection problem in dFC in this thesis can roughly be put into two categories. In the first one, predictive log-likelihood of held-out data is evaluated (Nielsen, Schmidt, et al., 2018), which can be applied to any data without the need for external variables (such a disease status or behavior). We stress that this does not necessarily promote interpretable models often sought in neuroimaging, i.e. the model that fits and generalizes best does not always offer insights into the mechanism that generated the data. Characterizing the data better does not necessarily mean that we understand the brain better.

In the other model selection category, the use of external information to select between dFC models is utilized. For the resting-state data and dFC models we have analysed (Nielsen, Levin-Schwartz, et al., 2018; Nielsen, Vidaurre, et al., 2018), we observed an inability to better discriminate subject groups compared to simple static modeling assumptions and inability to characterize higher-order behavioral traits. This could be due to not accounting properly for the noise process and spectral content in the data (Laumann et al., 2016; Nielsen, Schmidt, et al., 2018), leading to FC state estimates that are driven by these nuisance sources and are thus not relevant for the prediction at hand. This questions the desirability of the resting-state as an experimental paradigm. We do believe that the brain is dynamic in the sense that different cognitive and sensory processes require different functional network configurations. However, the important questions then are at what time-scale is the brain "dynamic“ and can we measure that with fMRI? Greene et al., 2018 analysed the prediction of different behavioural variables using (static) functional connectivity (FC) and found that models built on features from task-based FC outperformed the resting-state based models in the prediction of fluid intelligence. Greene et al., 2018 conclude, at least for the prediction of behavioural variables, that this motivates a paradigm-shift from rest to task-based FC.

Recent papers have investigated the utility of dFC models on task fMRI. Gonzalez-Castillo et al., 2015 used a SWKM to in an unsupervised manner extract a state sequence from an experimental paradigm with four different tasks. The state sequences extracted corresponded very well to the experimental conditions thus affirming the dFC states can have a meaningful interpretation in the right context. Furthermore, it has been show in Taghia et al., 2018 that particular dFC states are associated with the performance in a working memory task.

Since the first observations of dFC in resting-state fMRI a number of papers (Handwerker et al., 2012; Zalesky and Breakspear, 2015; Laumann et al., 2016; Liégeois et al., 2017; R. L. Miller et al., 2017) have dealt with the construction of an appropriate *null-model* for dFC. In frequentist hypothesis testing, the null-model is the one sampled to generate the null-data that will be compared to actual data. Now, it can be established if the observed data is too extreme for the null-model (and the null hypothesis can be rejected). For this to be useful,

the null-model must mimick many of the same properties observed in the real data, such as the noise process. In this thesis, we have taken a complementary approach rooted in the Bayesian framework in which candidate models are evaluated in their predictive performance on held-out data. A good practice in such a scenario is to have baseline model to asses the relative performance of the models of interest. The baseline model can thus act as a "null-model“ in the Bayesian framework given that it contains proper assumptions about the data at hand (as in the frequentist null-model). For a perspective on frequentist vs. Bayesian methods in a neuroscience context see Bzdok and Yeo, 2017.

In general, there are many theoretical and practical merits of Bayesian modeling. However as mentioned, these merits come at the price of making the inference of the parameters harder, both in terms of reliability and scalability. In the variational Bayesian (VB) framework, the inference is made tractable by updating the moments of a simpler distribution, $Q(\theta)$, to make $Q(\theta)$ "close“ to the true posterior. The inference can in that case get stuck in local minima, i.e. a hilltop where the gradient of the cost-function is zero, which calls for multiple restarts of the algorithm to test the reliability. The evaluation of the likelihood function can furthermore hamper the scalability of VB, which has been addressed by sampling noisy estimates of the gradient (Hoffman et al., 2013). Markov chain Monte Carlo (MCMC) methods have been shown to better escape local minima (Bishop, 2006) given enough samples in the chain. The problem then becomes diagnosing when we have "enough“ samples (Gelman et al., 2014). Scalability issues also exist for MCMC methods, which have been addressed similarly to scalable VB by subsampling approaches (Scott et al., 2016).

This thesis has displayed the utility of Bayesian modeling of dFC and furthermore the need for quantitative evaluation of models. Our work has illustrated the impact that different modeling assumptions have on the interpretation of brain dynamics and that there is a substantial variability not explicitly due to the experimental design or cognitive processing. This calls for better characterizations (or models) of the different forms of variability in data, i.e. subjects, sessions, task and noise confounds. Finally, we must decide what the ultimate goal is when modeling dFC in fMRI. Since all models are wrong (according to George Box), we must carefully decide what the emphasis of the dFC model should be in order to reach our ultimate goal, whether it's to characterize functional brain organization or prediction of cognitive traits and disease status.

Appendix A

Papers

A.1 Nonparametric Modeling of Dynamic Functional Connectivity in fMRI Data

Nielsen, Søren F V, Kristoffer H Madsen, Rasmus Røge, Mikkel N Schmidt, and Morten Mørup (2016). "Nonparametric Modeling of Dynamic Functional Connectivity in fMRI Data". In: *NIPS 2015 Workshop on Machine Learning and Interpretation in Neuroimaging*. Montreal, Canada: arxiv.org.

Nonparametric Modeling of Dynamic Functional Connectivity in fMRI Data

Søren F. V. Nielsen[1], Kristoffer H. Madsen[2], Rasmus Røge[1], Mikkel N. Schmidt[1], and Morten Mørup[1]

[1] Section for Cognitive Systems, DTU Compute, Technical University of Denmark
Richard Petersens Plads, Building 324, DK-2800 Kgs. Lyngby
sfvn@dtu.dk,
[2] Danish Research Centre for Magnetic Resonance, Section 714
Copenhagen University, Hospital Hvidovre, Kettegaard Allé 30, DK-2650 Hvidovre

Abstract. Dynamic functional connectivity (FC) has in recent years become a topic of interest in the neuroimaging community. Several models and methods exist for both functional magnetic resonance imaging (fMRI) and electroencephalography (EEG), and the results point towards the conclusion that FC exhibits dynamic changes. The existing approaches modeling dynamic connectivity have primarily been based on time-windowing the data and k-means clustering. We propose a nonparametric generative model for dynamic FC in fMRI that does not rely on specifying window lengths and number of dynamic states. Rooted in Bayesian statistical modeling we use the predictive likelihood to investigate if the model can discriminate between a motor task and rest both within and across subjects. We further investigate what drives dynamic states using the model on the entire data collated across subjects and task/rest. We find that the number of states extracted are driven by subject variability and preprocessing differences while the individual states are almost purely defined by either task or rest. This questions how we in general interpret dynamic FC and points to the need for more research on what drives dynamic FC.

Keywords: dynamic functional connectivity, Bayesian nonparametric modeling, hidden Markov modeling, Wishart mixture modeling, predictive likelihood

1 Introduction

The invention of functional magnetic resonance imaging (fMRI) paved the way for non-invasive studies of neuronal activity in the human brain. Especially the correlation of activity between brain regions, *functional connectivity* (FC), has been of interest for over a decade. In recent years the term *dynamic* functional connectivity has emerged in the field trying to explain temporal changes in the FC pattern [1,2,3]. Intuitively this makes sense since we expect that the interaction between segregated brain regions changes. These changes can be due to a number of factors such as the experimental design (task and resting state),

non-experimental physical factors (fatigue, caffeine intake) and even neurological disorders [4]. An important aim of modeling FC as a dynamic concept is to run the models in a fully-unsupervised setting to achieve greater knowledge of how the brain functions and even extract biomarkers for diseases [5]. One very prominent approach to analysing dynamic FC was presented in [6]. Here windowed covariance matrices was extracted from a group-ICA representation of 405 healthy subjects resting state fMRI data. The upper triangular part of each extracted covariance matrix was then stacked into a vector yielding a vector space representation of the covariance structure for each window and subject. These vectors were then clustered using K-means clustering and the number of clusters was chosen using the Elbow-criterion. The results indicated that the resting state paradigm exhibits a temporally dynamical structure and that some of the FC patterns extracted diverge from the classical stationary results (such as the default mode network). In order to validate that the FC patterns (or states) extracted can be used to characterize resting state, we must test that the models find meaningful results in a setting where we have ground truth information available. [7] constructed an experiment where the participants were scanned while being asked to do different tasks in one continuous scan. They found that using windowed covariance matrices and k-means clustering (similar to [6]) the cluster centroids could be used to explain the tasks carried out. But the number of dynamic states k (fixed to the number of tasks) and window length had to be pre-specified whereas the model was not evaluated on independent test data.

In this paper we propose an extension of the infinite hidden Markov model (IHMM) [8] tailored for the modeling of dynamic FC states in fMRI. We furthermore present a predictive likelihood framework for validating that the extracted structure can be used to characterize a motor task from rest. Using the IHMM circumvents having to specify the number of states in the model and window length. Instead the model can be viewed as an adaptive windowing method, where states can persist on different state specific time scales and the number of states learned as part of the inference. We will use this framework to investigate what drives dynamic FC.

2 Methods

The IHMM-Wishart model: A commonly used representation of FC is to represent the connectivity between areas of the brain by the covariance matrix. In a dynamic setting we model each state as having separate covariance matrices. We thus model fMRI in terms of a latent sequence of states, z_t for $t = 1..T$ that for each time point generates a brain image according to the state specific covariance matrix $\mathbf{x}_t \sim \mathcal{N}(\mathbf{0}, \sigma_t^2 \boldsymbol{\Sigma}^{(z_t)})$, where σ_t^2 defines the magnitude of the state specific covariance structure $\boldsymbol{\Sigma}^{(k)}$ at time t allowing states to be invariant to data magnitude but defined in terms of the connectivity profile. For the modeling of transitions between states we use the infinite hidden Markov model (IHMM), first proposed in [8] and further developed in [9,10]. Completing the IHMM with a Gaussian distribution on the observed data and an inverse Wishart prior on

the covariance structure, similar to the Wishart mixture modeling considered in [11,12] and the infinite Gaussian mixture model of [13], we arrive at the following generative model,

$$\boldsymbol{\beta} \sim \text{GEM}(\gamma), \tag{1}$$

$$\boldsymbol{\pi}^{(k)}|\boldsymbol{\beta} \sim \text{DP}(\alpha, \boldsymbol{\beta}), \tag{2}$$

$$z_t|z_{t-1} \sim \text{Multinomial}(\boldsymbol{\pi}^{(z_{t-1})}), \tag{3}$$

$$\boldsymbol{\Sigma}^{(k)} \sim \mathcal{W}^{-1}(\eta\boldsymbol{\Sigma}_0, v_0), \tag{4}$$

$$\mathbf{x}_t \sim \mathcal{N}(\mathbf{0}, \sigma_t^2\boldsymbol{\Sigma}^{(z_t)}). \tag{5}$$

GEM is the stick-breaking construction (cf. [14,15]), DP is the Dirichlet process, γ and α are positive hyper-parameters controlling the state sequence, $\boldsymbol{\pi}^{(k)}$ denotes the k'th row of the transition matrix $\boldsymbol{\pi}$, η a is (positive) scale parameter, $\boldsymbol{\Sigma}_0$ is a $p \times p$ matrix (covariance prior), and $\mathcal{W}^{-1}(\boldsymbol{\Sigma}_0, v_0)$ denotes the inverse Wishart distribution, and σ_t^2 is the time specific covariance scaling. Viewing this in light of fMRI and FC (cf. also [16,12]), each state can further be represented by a covariance matrix defining the FC where each state is invariant to magnitude of the FC due to the time-specific scaling parameter σ_t^2. We learn this parameter in the inference procedure (cf. next section) using a vague improper $1/\mathcal{X}$-prior. $\boldsymbol{\Sigma}_0$ in the inverse Wishart prior plays the role of the default connectivity. In the experiments we estimate $\boldsymbol{\Sigma}_0$ from a separate resting state fMRI scan. The parameter η is the scaling or level of this default connectivity, which is learned during the inference. The degrees of freedom, v_0, set to the number of dimensions p to make the inference well posed.

Inference: For inference we use Markov chain Monte Carlo (MCMC). Due to conjugacy we can analytically integrate the state specific covariances $\boldsymbol{\Sigma}^{(k)}$ and infer the state sequence $\boldsymbol{z}$ using Gibbs sampling as described in [10] and split-merge sampling [17,18] which has been demonstrated to improved mixing and convergence properties for non-parametric Bayesian mixture models. We learn η and σ_t^2 by Metropolis-Hastings with proposal distribution, $\eta^* = \exp(\ln\eta + z), z \sim \mathcal{N}(0, \sigma = 0.1)$ (and similarly for σ_t^2) imposing an improper and uninformative $1/\mathcal{X}$-prior on both variables. The implementation of the model was done in MATLAB building on top of Juergen Van Gael's iHMM-toolbox [19]. We used the code available online[3] for sampling state-sequence relevant hyperparameters, where vague Gamma priors are placed on α and γ and inferred via. an auxiliary variable Gibbs sampler [20]. The predictive likelihood can be calculated using a modified Viterbi algorithm [21], and analytically integrating out σ_t^2 and using parameter posterior samples from the inference.

3 Results

We used fMRI data from a population of 29 subjects, that carried out a simple finger-tapping experiment (denoted motor) [22] as well as a resting state scan

[3] `http://mloss.org/software/view/205/`

[23]. Preprocessing included normalization to MNI space and wavelet despiking [24]. We did a group ICA [25] using the GIFT toolbox[4] into 20 components using the maximum likelihood ERBM algorithm [26] with default settings. We discarded six components by visual inspection of the spatial maps representing common noise confounds. We split each subjects motor and resting-state data into a training (116 images) and a test set (120 images). Each training and test set was individually corrected for motion- and respiratory effects [27], reference signal from cerebrospinal fluid (lateral ventricles) and white matter along with high-pass filtering (1/128 s cut-off) in a single regression step. The resting state scan was twice as long as the motor experiment so the second half of the resting state scan was removed and used for estimating the prior covariance $\boldsymbol{\Sigma}_0$. For each of the 29 subjects training sets available (motor and rest) we ran ten IHMM-Wishart models and ten constant Wishart models, i.e. an IHMM-Wishart forced to be in one state.

First of all, we investigate if a model trained on the motor task (denoted a motor-model) predicts better on the motor-test set relative to a model trained on the resting state (denoted a rest-model). Second of all, we can test if predictive performance increases using a dynamic model, compared to a one-state constant model. For the motor data, we calculated the predictive likelihood for all ten runs, and made a paired t-test to evaluate if the predictive performance was better by the motor-model or the rest-model. The same was done for the resting state data. We saw on test data that within all 29 subjects, motor and rest was distinguishable with an IHMM-Wishart model. We furthermore tested if there was a significant difference between using a dynamic model vs. a static model. Except for one subject, this difference was not significant, which can be explained by the IHMM-Wishart finding almost exclusively one state on both motor and rest.

The above analysis can be extended to group level, i.e. to investigate if a subjects task and resting state data can be distinguished by models trained on other subjects. We looked at 4 hypothesis tests, 1) that the motor- and rest-model predict equally well on motor data, 2) that the rest- and motor-model predict equally well on resting state data, 3) that the motor-model is equal to using a one state constant model on motor data and 4) that the rest-model is equal to using a one state constant model on the resting state data. For motor hypothesis 1 was rejected on average 26.5 ± 2 out of 28 times for a subject, and for resting-state hypothesis 2 was rejected on average 26.1 ± 2 out of 28 times. Similarly as the within-subject analysis, very few dynamic models perform better on testing data than the one-state counterpart models. Hypothesis 3) and 4) could only be rejected 0.31 ± 0.5 and 0.24 ± 0.6 times respectively.

To investigate dynamic FC at group level, we collated all subjects motor and resting state data together and ran the IHMM-Wishart fully unsupervised on the whole data set. In the modeling, states are shared across subjects and tasks but we handle discontinuities in the data by restarting the chain at each block that has been preprocessed individually (4 per subject). We ran the IHMM-Wishart

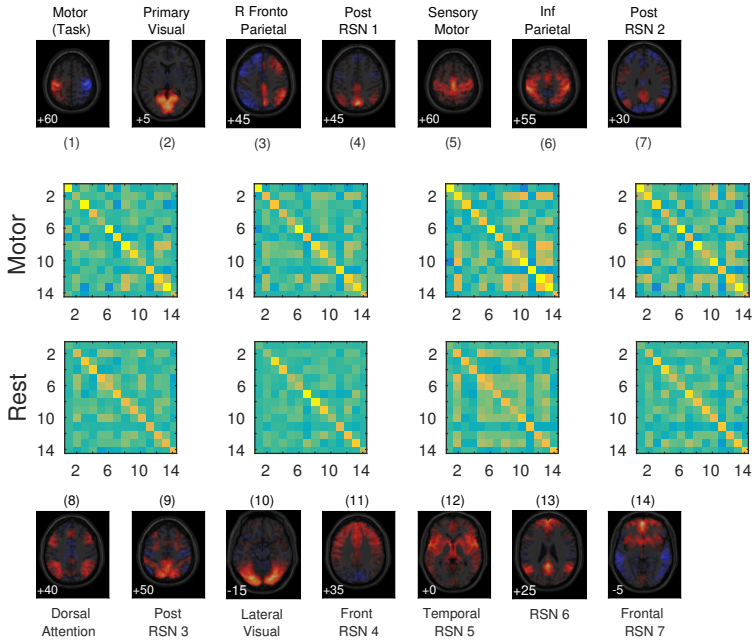
[4] `http://mialab.mrn.org/software/gift/`

five times, and the sample with highest posterior probability was taken out for further analysis. The results can be seen in figure 1. In figure 1a we report the four most populated states' covariance matrices in both motor and resting-state along with the ICA components (IC). We note IC1 that well corresponds to the motor task is active in all the motor-states, and has been down-weighted in the resting-states. Inspecting the two tasks, the average number of states pr. subject in the motor-task was 3.28 ± 0.21 and for the resting state 3.60 ± 0.27. Furthermore, the states extracted are generally pertaining almost perfectly to either motor or rest (cf. figure 1b). However, from figure 1c we see that the average mutual information (MI) over posterior samples with state sequences describing the subjects, task and preprocessing shows that subjects and preprocessing drives the dynamics more than the tasks themselves.

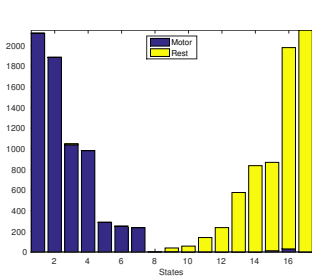
4 Discussion

The IHMM-Wishart can distinguish between motor and rest on test data from the same subject. One confound is that the two experiments are recorded separately meaning that our ability to distinguish between motor and resting-state test data probably is caused by training and test set being recorded close in time. Since the group analysis shows that models trained on a subject can well distinguish between motor and rest of others, this confound is not present in this setting. This means that the IHMM-Wishart is able to well extract states that characterize motor from rest. In this analysis the model only found support for a single state in most of these single subject analyses carried out. The number of states found by the IHMM-Wishart is influenced though by the number of data points available and dimensionality, and this could influence the 'absence' of dynamics found due to the size of the training set (116 images). Collating data together from all subjects, we ran our model on the entire data set and saw each state were clearly pertaining to mainly either motor or rest. We see though that preprocessing has an impact on the states extracted - in an information theoretic perspective, the dynamics found are closer related to preprocessing than both to subjects and task (cf. figure 1c).

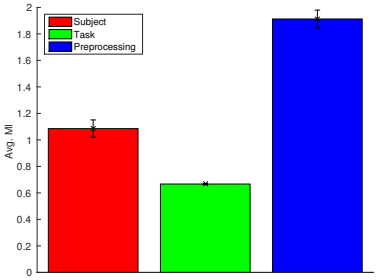
Our predictive framework demonstrates that the IHMM-Wishart can be used to characterize task vs. rest. This is a very controlled setting and should also be possible to achieve for simpler models than the one we are proposing here but evidences the utility of the IHMM-Wishart in characterizing fMRI data. Importantly, the IHMM-Wishart enables us to infer the number of dynamic states from data which for the collated data turned out to be *more* than just two states, i.e. one for rest and one for motor. We here observed multiple task specific dynamic states driven by subject variability and preprocessing, which is aligned with the conclusion in [12]. This tells us that care must be taken when interpreting dynamics at a group level as preprocessing even of blocks of data from the same subject evidences the presence of multiple dynamic functional states.



(a) Covariance estimate for each of the most populated states in the motor task (top row) and resting state (bottom row), along with the ICA component maps.



(b) A bar plot displaying how frequently each state is populated. The states have been sorted by the difference in frequency in motor and rest-state.



(c) Average mutual information (MI) over samples with three state sequences, 1) subject (red), 2) task (green) and 3) preprocessing (blue)

Fig. 1: We collated all subjects motor and resting state experiments together and ran 5 chains of the IHMM-Wishart model. We show how the different states are populated, and investigate what drives the dynamics; subjects, tasks or preprocessing, for the sample with highest posterior probability.

Acknowledgements

This work was supported by the Lundbeck Foundation, grant no. R105-9813.

References

1. R Matthew Hutchison, Thilo Womelsdorf, Elena A Allen, Peter A Bandettini, Vince D Calhoun, Maurizio Corbetta, Stefania Della Penna, Jeff H Duyn, Gary H Glover, Javier Gonzalez-Castillo, et al. Dynamic functional connectivity: promise, issues, and interpretations. *Neuroimage*, 80:360–378, 2013.
2. Andrew Zalesky, Alex Fornito, Luca Cocchi, Leonardo L Gollo, and Michael Breakspear. Time-resolved resting-state brain networks. *Proceedings of the National Academy of Sciences*, 111(28):10341–10346, 2014.
3. Vince D Calhoun, Robyn Miller, Godfrey Pearlson, and Tulay Adalı. The chronnectome: time-varying connectivity networks as the next frontier in fMRI data discovery. *Neuron*, 84(2):262–274, 22 October 2014.
4. Vince D Calhoun, Tom Eichele, and Godfrey Pearlson. Functional brain networks in schizophrenia: a review. *Frontiers in human neuroscience*, 3, 2009.
5. Archana Venkataraman, Thomas J Whitford, Carl-Fredrik Westin, Polina Golland, and Marek Kubicki. Whole brain resting state functional connectivity abnormalities in schizophrenia. *Schizophrenia research*, 139(1):7–12, 2012.
6. Elena A Allen, Eswar Damaraju, Sergey M Plis, Erik B Erhardt, Tom Eichele, and Vince D Calhoun. Tracking whole-brain connectivity dynamics in the resting state. *Cerebral cortex*, page bhs352, 2012.
7. Javier Gonzalez-Castillo, Colin W Hoy, Daniel A Handwerker, Meghan E Robinson, Laura C Buchanan, Ziad S Saad, and Peter A Bandettini. Tracking ongoing cognition in individuals using brief, whole-brain functional connectivity patterns. *Proc. Natl. Acad. Sci. U. S. A.*, 112(28):8762–8767, 14 July 2015.
8. Matthew J Beal, Zoubin Ghahramani, and Carl E Rasmussen. The infinite hidden markov model. In *Advances in neural information processing systems*, pages 577–584, 2001.
9. Alan S Willsky, Erik B Sudderth, Michael I Jordan, and Emily B Fox. Nonparametric bayesian learning of switching linear dynamical systems. In *Advances in Neural Information Processing Systems*, pages 457–464, 2009.
10. Jurgen Van Gael. *Bayesian Nonparametric Hidden Markov Models*. PhD thesis, University of Cambridge, 2012.
11. Sullivan Hidot and Christophe Saint-Jean. An expectation–maximization algorithm for the wishart mixture model: Application to movement clustering. *Pattern Recognition Letters*, 31(14):2318–2324, 2010.
12. Josefine Korzen, Kristoffer H Madsen, and Morten Mørup. Quantifying temporal states in rs-fmri data using bayesian nonparametrics. In *HBM 2014, Poster Number: 1726*. Organization for Human Brain Mapping, 2014.
13. Carl Edward Rasmussen. The infinite gaussian mixture model. In *NIPS*, volume 12, pages 554–560, 1999.
14. J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.
15. Jim Pitman. Poisson–dirichlet and gem invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability & Computing*, 11(05):501–514, 2002.

16. Srikanth Ryali, Kaustubh Supekar, Tianwen Chen, Weidong Cai, and Vinod Menon. A variational bayes hidden markov model for discovering dynamical functional brain networks. In *HBM 2015, Abstract 3749*. Organization for Human Brain Mapping, 2015.
17. Sonia Jain and Radford M Neal. A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1), 2004.
18. Michael C Hughes, Erik B Sudderth, and Emily B Fox. Effective split-merge monte carlo methods for nonparametric models of sequential data. In *Advances in Neural Information Processing Systems*, pages 1295–1303, 2012.
19. J. Van Gael. The infinite hidden markov model, 2010. `http://mloss.org/software/view/205/`.
20. Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476), 2006.
21. Andrew J Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2):260–269, 1967.
22. Peter M Rasmussen, Lars K Hansen, Kristoffer H Madsen, Nathan W Churchill, and Stephen C Strother. Model sparsity and brain pattern interpretation of classification models in neuroimaging. *Pattern Recognition*, 45(6):2085–2100, 2012.
23. Kasper Winther Andersen, Kristoffer H Madsen, Hartwig Roman Siebner, Mikkel N Schmidt, Morten Mørup, and Lars Kai Hansen. Non-parametric bayesian graph models reveal community structure in resting state fmri. *NeuroImage*, 100:301–315, 2014.
24. Ameera X Patel, Prantik Kundu, Mikail Rubinov, P Simon Jones, Petra E Vértes, Karen D Ersche, John Suckling, and Edward T Bullmore. A wavelet method for modeling and despiking motion artifacts from resting-state fMRI time series. *Neuroimage*, 95:287–304, 15 July 2014.
25. VD Calhoun, T Adali, GD Pearlson, and JJ Pekar. A method for making group inferences from functional mri data using independent component analysis. *Human brain mapping*, 14(3):140–151, 2001.
26. Xi-Lin Li and Tülay Adali. Blind spatiotemporal separation of second and/or higher-order correlated sources by entropy rate minimization. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 1934–1937. IEEE, 2010.
27. K J Friston, S Williams, R Howard, R S Frackowiak, and R Turner. Movement-related effects in fMRI time-series. *Magn. Reson. Med.*, 35(3):346–355, March 1996.

A.2 Modeling Dynamic Functional Connectivity using a Wishart Mixture Model

Nielsen, Søren F V, Kristoffer H Madsen, Mikkel N Schmidt, and Morten Mørup (2017). "Modeling dynamic functional connectivity using a wishart mixture model". In: *2017 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*. Toronto, Canada: IEEE, pp. 1–4. DOI: 10.1109/PRNI.2017.7981505.

Modeling Dynamic Functional Connectivity using a Wishart Mixture Model

Søren F.V. Nielsen*, Kristoffer H. Madsen*,†, Mikkel N. Schmidt* and Morten Mørup*
* DTU Compute, Technical University of Denmark, Denmark
†Danish Research Centre for Magnetic Resonance, Copenhagen University Hospital Hvidovre, Denmark
Corresponding email: sfvn@dtu.dk

***Abstract*—Dynamic functional connectivity (dFC) has recently become a popular way of tracking the temporal evolution of the brains functional integration. However, there does not seem to be a consensus on how to choose the complexity, i.e. number of brain states, and the time-scale of the dynamics, i.e. the window length. In this work we use the Wishart Mixture Model (WMM) as a probabilistic model for dFC based on variational inference. The framework admits arbitrary window lengths and number of dynamic components and includes the static one-component model as a special case. We exploit that the WMM framework provides model selection by quantifying models generalization to new data. We use this to quantify the number of states within a prespecified window length. We further propose a heuristic procedure for choosing the window length based on contrasting for each window length the predictive performance of dFC models to their static counterparts and choosing the window length having largest difference as most favorable for characterizing dFC. On synthetic data we find that generalizability is influenced by window length and signal-to-noise ratio. Too long windows cause dynamic states to be mixed together whereas short windows are more unstable and influenced by noise and we find that our heuristic correctly identifies an adequate level of complexity. On single subject resting state fMRI data we find that dynamic models generally outperform static models and using the proposed heuristic points to a window-length of around 30 seconds provides largest difference between the predictive likelihood of static and dynamic FC.**

I. INTRODUCTION

It is a well know fact that the brains way of integrating and segregating information changes over time and is perturbed by various cognitive tasks. In functional magnetic resonance imaging (fMRI) the functional brain "network" is often described using functional connectivity (FC) models, i.e. the correlation between segregated regions of interest, and these networks are known to change during task. In recent years studies of resting-state FC have shown to also exhibit dynamic properties indicating that FC during rest is non-stationary. Thus tracking temporal changes in FC during resting-state has become a popular research topic in recent years [1], [2], [3]. We see two main advantages of modeling the temporal changes in FC; first there has been some evidence that the dFC can be used to characterize different psychiatric disorders, such as PTSD [4] and schizophrenia [5]. Secondly, we hope by modeling dynamic functional connectivity (dFC) to gain a better understanding of the resting-state and the spontaneous changes in coupling between regions not associated by task activation [6].

Most models of dFC use the *sliding-window* approach [6], [7], where the assumption is that the FC is stable in subsegments of the data. Allen et. al. [6] applied this to a large cohort of healthy subjects where the extracted region-by-region covariance matrices from each window were clustered using the k-means algorithm. The 7 states extracted showed a highly non-stationary behavior where coupling in and to the default mode network notably varied over the states. We are though still faced with a number of problems in estimating reliable dFC patterns [8], [9], [10]. On one hand we face the problem of timescales, i.e. window length of dFC patterns is in most cases assumed to be known or set to some value based on the acquisition parameters in the experiment. Window-free methods exists, such as hidden Markov models in the context of microstates for EEG/MEG [11] and for dFC in fMRI [12], [13]. However, these models are more expressive and thus it becomes of importance to control the "over-characterization" in the training. Finally, in both microstate analysis and dFC models the complexity, i.e. the number of states to extract, is always a problem [13]. In k-means we have no natural way to choose the number of states, and thus heuristics such as the Gap-criterion is often used.

In this work we will use a Bayesian formulation of the Wishart mixture model (WMM) [14], [15], [16]. The Wishart distribution is defined as the distribution of the scatter matrix of zero-mean multivariate Gaussian data, and is thus a natural likelihood function for windowed functional connectivity. Whereas [14] used expectation-maximization (EM) and [16] Gibbs sampling we presently consider variational inference and use the WMM as a probabilistic analogy to the sliding-window k-means clustering approach. The probabilistic treatment will allow us to tap into features of the Bayesian modeling framework such as prediction. We will investigate a predictive likelihood framework to estimate the number of states in dFC problems, both in a synthetic setting and in resting-state fMRI data and propose a heuristic for choosing the window length.

*Søren F.V. Nielsen, Mikkel N. Schmidt and Morten Mørup were supported by Lundbeckfonden (fellowship grant R105-9813 to Morten Mørup). Kristoffer H. Madsen was supported by a Novo Nordisk Foundation Interdisciplinary Synergy Grant (NNF14OC0011413)

978-1-5386-3159-1/17/$31.00 ©2017 IEEE

II. METHODS

We first briefly present some notation. Let $\boldsymbol{x}_t \in \mathbb{R}^p$ be a zero-mean distributed signal at time point t. We consider data of L symmetric-positive semi-definite matrices C_ℓ of size $p \times p$ in which there are K clusters. In this paper we will use Gram matrices, i.e. $C_\ell = \sum_{t \in W_\ell} \boldsymbol{x}_t \boldsymbol{x}_t^T$, in which W_ℓ is the ℓ'th window set.

A. Bayesian Wishart Mixture Model

The Bayesian Wishart Mixture Model (WMM) for K states can be written in terms of the generative model,

$$\begin{aligned} \boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha}) \quad z_\ell \sim \mathrm{Cat}(\boldsymbol{\pi}) \quad \eta \sim \mathcal{G}^{-1}(a_0, b_0) \\ {\boldsymbol{\Sigma}^{(k)}}^{-1} \sim \mathcal{W}(\eta \mathbf{I}_p, \nu_0) \quad \mathbf{C}_\ell \sim \mathcal{W}(\boldsymbol{\Sigma}^{(z_\ell)}, \nu_\ell), \end{aligned} \tag{1}$$

in which $\boldsymbol{\pi}$ is a vector of length K containing the proportions of the states, $\boldsymbol{\alpha}$ is the prior vector of length K for the Dirichlet distribution, z_ℓ is the categorical (hard)assignment of window ℓ, η is the prior on the "scale" of the cluster centres, ${\boldsymbol{\Sigma}^{(k)}}^{-1}$ is the k'th cluster centres inverted Gram matrix, ν_0 is the degrees of freedom for each cluster centre and ν_ℓ is the degrees of freedom for the ℓ'th window. The prior on the cluster centres, ${\boldsymbol{\Sigma}^{(k)}}^{-1}$, and the parameter η is mostly chosen for convenience in the inference procedure and makes all updates closed form. The η parameter works as a regularizer on the cluster centres. This becomes very important in high dimensions and a low number of data points, as is the case in most fMRI scenarios. As for the degrees of freedom for each window, ν_ℓ, we propose summing the window-function, i.e. yielding ν_ℓ equal to the window length for the box-car window.

B. Variational Bayes

As with many Bayesian models evaluating the posterior, $p(\boldsymbol{\theta}|\boldsymbol{X})$, is intractable due to the model evidence term, $p(\boldsymbol{X})$, obtained from Bayes rule. We therefore turn to the variational Bayesian (VB) framework to approximate the posterior. In VB the goal is to find a distribution, $Q(\boldsymbol{\theta})$, which is "close" in the Kullback-Leibler (KL) divergence to the posterior and has a simpler form such that inference becomes tractable. We choose to use the well-known mean-field approximation in which the distribution of each parameter is assumed to factorize. Minimizing the KL-divergence between the intractable posterior and the Q-distribution is equivalent to maximizing the evidence lower-bound (ELBO), which can be formulated as,

$$\begin{aligned} \mathcal{L}(\mathbf{C}, \theta) = \langle \log p(\mathbf{C}|\boldsymbol{\Sigma}^{-1}, \mathbf{z}, \boldsymbol{\pi}) \rangle + \langle \log p(\boldsymbol{\Sigma}^{-1}) \rangle + \langle \log p(\eta) \rangle \\ + \langle \log p(\mathbf{z}|\boldsymbol{\pi}) \rangle + \langle \log p(\boldsymbol{\pi}) \rangle - \langle \log Q(\boldsymbol{\Sigma}^{-1}) \rangle \\ - \langle \log Q(\eta) \rangle - \langle \log Q(\mathbf{z}) \rangle - \langle \log Q(\boldsymbol{\pi}) \rangle, \end{aligned} \tag{2}$$

in which $\mathcal{L}$ is the ELBO, $\mathbf{C}$ is the collection of all the windowed scatter matrices, $\boldsymbol{\Sigma}^{-1}$ is the collection of all ${\boldsymbol{\Sigma}^{(k)}}^{-1}$, $\mathbf{z}$ is a vector of length L containing all z_ℓ, $\langle \cdot \rangle$ denotes expectation wrt. the Q-distribution, and θ is the collection of all parameters in the model. Now we maximize the ELBO using coordinate ascend variational inference (CAVI), which results in calculating the sufficient statistics of each factor (all of which are closed-form) while keeping the others fixed in a cyclic fashion. The Q-distributions along with the update rules have the form,

$$Q(\boldsymbol{\Sigma}^{-1}) = \prod_k \mathcal{W}({\boldsymbol{\Sigma}^{(k)}}^{-1} | \boldsymbol{\Omega}^{(k)}, v_k) \tag{3}$$

$$\boldsymbol{\Omega}^{(k)} = \left(\boldsymbol{\Sigma}_0^{-1} + \sum_\ell \langle z_{\ell k} \rangle \mathbf{C}_\ell \right)^{-1}, v_k = \nu_0 + \sum_\ell \langle z_{\ell k} \rangle \nu_\ell$$

$$Q(\eta) = \mathcal{G}^{-1}(\eta | \tilde{a}, \tilde{b}) \tag{4}$$

$$\tilde{a} = a_0 + \frac{\nu_0 p K}{2}, \tilde{b} = b_0 + \frac{1}{2} \sum_k \mathrm{tr}(\langle {\boldsymbol{\Sigma}^{(k)}}^{-1} \rangle)$$

$$Q(\mathbf{z}) = \prod_\ell \mathrm{Cat}(\mathbf{z}_\ell | \mathbf{r}_\ell) \tag{5}$$

$$\begin{aligned} \tilde{r}_{\ell k} = \frac{\nu_\ell - p - 1}{2} \ln |\mathbf{C}_\ell| - \frac{\nu_\ell}{2} \ln(2) - \ln \Gamma_p(\frac{\nu_\ell}{2}) \\ - \frac{\nu_\ell}{2} \langle \ln |\boldsymbol{\Sigma}^{(k)}| \rangle - \frac{1}{2} \mathrm{tr}(\langle {\boldsymbol{\Sigma}^{(k)}}^{-1} \rangle \mathbf{C}_\ell) + \langle \ln \pi_k \rangle \\ r_{\ell k} = \frac{\exp(\tilde{r}_{\ell k})}{\sum_{k'} \exp(\tilde{r}_{\ell k'})} \end{aligned}$$

$$Q(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\mathbf{a}), \quad a_k = \sum_\ell \langle z_{\ell k} \rangle + \alpha_k \tag{6}$$

In all experiments the following parameters were fixed: $\boldsymbol{\alpha} = [1, 1, ..., 1]$ and $\nu_0 = p$. Note that if $\boldsymbol{C}_\ell$ does not have full rank some terms in (5) cannot be computed. These terms are however constant over k and can thus be ignored.

C. Predictive Likelihood and Bayes Factors

It is a well-known fact that VB is prone to underestimating the posterior variance [17], and therefore we do not usually rely only on the ELBO to do model selection. Thus, we need a more conservative measure that promotes generalizable models. We thus turn to predictive likelihood on previously unseen data, $\boldsymbol{C}^*$, which is dependent on the choice of the number of states K, i.e.

$$p(\boldsymbol{C}^*|\boldsymbol{C})_K = \int p(\boldsymbol{C}^*|\boldsymbol{\theta})_K p(\boldsymbol{\theta}|\boldsymbol{C})_K d\boldsymbol{\theta} \tag{7}$$

Since we do not have access to the true posterior, we use the approximation $Q(\boldsymbol{\theta})_K \approx p(\boldsymbol{\theta}|\boldsymbol{C})_K$, and due to the structure of the likelihood and the Q-distribution the approximation can be calculated analytically. In the following we will run the inference for a different number of states and calculate the predictive log Bayes factor, BF_k, towards the static model (with $K = 1$),

$$BF_k = \log p(\boldsymbol{C}^*|\boldsymbol{C})_k - \log p(\boldsymbol{C}^*|\boldsymbol{C})_1 \tag{8}$$

D. Generating Synthetic Data

To investigate the models capabilities and to validate our implementation we ran a number of synthetic experiments. In the following sections we will refer to data as being "synthetic" meaning that data generated by the following process. First, we generate K random upper triangular matrices $\mathbf{R}_k$ of dimension $p \times p$ by drawing each non-zero element of $\mathbf{R}_k$ from a standard

normal distribution $\mathcal{N}(0,1)$. In all of our experiments the number of states K was set equal to three. Now we fix a "true" window length, w_α, and for each subsegment of the synthetic data first draw a random state (i.e. a number from $1..K$) and then w_α observations from $\mathcal{N}(\mathbf{0}_p, \mathbf{R}_k^T\mathbf{R}_k)$. This yields a data matrix, $\boldsymbol{X}_{\text{signal}}$, of size $p \times T$. Finally, we generate white noise, $\boldsymbol{X}_{\text{noise}} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$, and create a linear combination of the data and noise to control the signal-to-noise (SNR) ratio, i.e. $\boldsymbol{X} = \gamma\boldsymbol{X}_{\text{signal}} + (1-\gamma)\boldsymbol{X}_{\text{noise}}$. We do this process independently for the training and test data.

E. Resting State Data

We analyze the single subject dataset from [18] containing resting-state fMRI sessions collected over a period of 18 months. Using SPM 12[1] revision 6685, we applied the following preprocessing steps to sessions 014-104 . All resting-state sessions were coregistred to the first image of session 014. We jointly corrected all sessions for motion artefacts using a rigid-body transformation towards the mean volume. An anatomical image from session 012 (T1 weighted) was coregistred to the functional space and a tissue probability map for grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF) extracted using the standard map from SPM12. Next we applied bandpass filtering [0.009-0.08] Hz, nuisance regression (motion parameters, eroded CSF and WM masks) and wavelet despiking [19]. The images were then resliced (due to a change in the number of slices after session 027) to the first session and smoothed using a FWHM 5mm Gaussian kernel. After preprocessing we ran a group ICA using the GIFT software[2] version 4.0a using 85 components, the ERBM algorithm and otherwise default settings.

III. RESULTS

A. Synthetic Data

To investigate the influence of window length and SNR on the predictive framework, we conduct a synthetic study with the following fixed parameters: $w_\alpha = 10$, $p = 10$, $T = 10000$ and fixed $\eta^{-1} = 1e-4$ in the model. We restarted each model inference 10 times and varied the number of states $K = 1..10$. Furthermore, we repeated the data generation process 10 times, and the mean BF_k over data sets (including standard deviation as error bars) can be seen in figure 1. In the noise-less case (top-left, $\gamma = 1$), we note that when the window-length is sufficiently small ($w \leq 10$) the model estimates the true number of states $K = 3$. However, we see an overestimation when the window-length becomes larger than w_α. This can be explained by the data being very inhomogeneous, and longer window lengths will be penalized more due to the mixing of different states within a window. Thus the models with longer window lengths need more states to explain the data. This effect gradually disappears as we decrease the SNR. One thing to note for all SNR levels is that the window lengths that are shorter than or equal to w_α seem to have a larger

[1]http://www.fil.ion.ucl.ac.uk/spm/

[2]http://mialab.mrn.org/software/gift/index.html

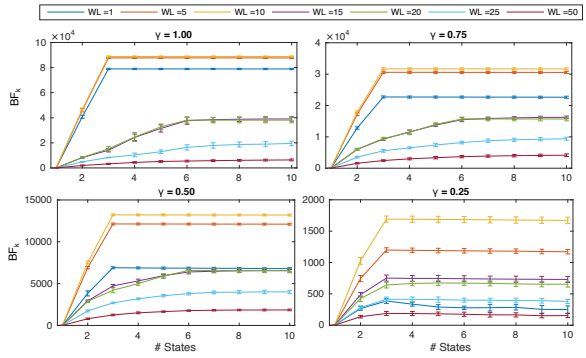


Fig. 1: Synthetic experiments for the influence of window length. We generated 10 synthetic data sets with states that had a "true" window length of $w_\alpha = 10$ for different SNR levels $\gamma = [1, 0.75, 0.5, 0.25]$. Then we ran our predictive likelihood framework for different window lengths, and we plot the mean and standard deviation of the BF_k over data sets.

increase in Bayes factor from the static $K = 1$ state model to the best predictive performance saturating at the $K = 3$ state model when compared to the longer window lengths. However, when we reach a certain noise level (in this case $\gamma = 0.25$) the shortest window length, $\text{WL} = 1$ has a flatter curve. Thus we find ourselves in a trade-off between window length and SNR; we want to make the window length low enough in order to not mix states together, but on the other hand not too low such that the estimation becomes unstable.

B. Single Subject Resting-State fMRI

To test the predictive framework on real data, we analysed a single subject resting-state fMRI data from [18]. Due to computational complexity, we ran the inference on 10 random subsets of data, each containing 5 sessions, and then calculated the predictive likelihood on the remaining sessions. It should be noted that some of the training sessions were in multiple training subsets. Each inference was restarted 10 times. Furthermore, we fixed η^{-1} during inference but varied its value in the range $[10^{-5}, 10^5]$ (sampled at ten points equidistantly in the log-domain). We choose the η^{-1}-value that yields the best predictive likelihood. The mean BF_k (over subsets) as a function of the number of states in the model is shown in figure 2. We see that the lowest window length has an almost flat curve, meaning that all number of states is equally likely, indicating that the window length is too short. The window length having the highest contrast between static and dynamic modeling, thus having the greatest increase in log Bayes factor before hitting a plateau is $\text{WL} = 25$ (i.e., around 30 seconds).

IV. DISCUSSION & CONCLUSION

We have proposed the Wishart mixture model (WMM) as a probabilistic extension of windowed k-means, to model dynamic functional connectivity in fMRI. As a way to quantify the number of states best accounting for dFC we use the

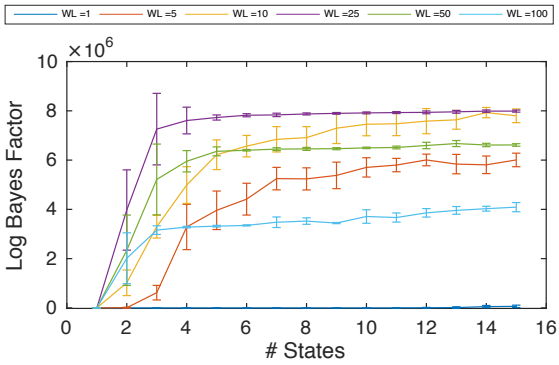


Fig. 2: Experiment on resting-state fMRI data from a single subject. We randomly sampled 5 resting-state sessions, trained the VB-WMM model with a different number of states $K = 1..15$ and calculated the predictive likelihood on the remaining sessions. We repeated this process 10 times with new random training subsets. In the figure the mean BF_k (over training splits) on the held out sessions is plotted along with one standard deviation as errorbars. The entire analysis was done for different window lengths (WL) in TRs, $\mathrm{WL} = [1, 5, 10, 25, 50, 100]$.

predictive likelihood. We further proposed a heuristic based on contrasting the predictive likelihood of dFC to the predictive likelihood of the corresponding static model containing only one state in order to quantify a suitable window length for characterizing dFC. On synthetic data we found that this heuristic correctly indicated the correct level of complexity. On real single subject resting state data we found support for dynamic modeling for all considered window lengths except ($WL = 1$ where the static model ($K = 1$) was not outperformed by dynamic models ($K > 1$)) and using the heuristic of highest contrast in predictive likelihood between static and dynamic modeling we found $WL = 30$ most suitable for characterizing dFC.

We would like to emphasize that the proposed procedure for quantifying window-length is a heuristic that we find useful to quantify trade-offs between SNR and issues mixing dynamic states but that predictive performance using different window lengths cannot be directly compared as they are based on test data having different properties. It should also be noted that as we increase the window length, we have fewer and fewer data points for training the WMM, which could influence the results. One could look into using overlapping windows to mitigate the effect of mixing states together, which is an avenue to pursue in future work. Also, in this work we have used Gram matrices to "stay true" to the likelihood function we are using in the WMM. However, there might be differences in using covariance matrices or even correlation matrices, which should be investigated further. In the real data there could be a pitfall caused by noisy ICA components, i.e. we have not done any post-selection. However, if states were driven by noise components they are not likely to generalize well to new sessions, and the predictive likelihood should in theory take care of this. However, to really confirm this would require an in-depth analysis. Thus, the proposed heuristic needs to be further validated. In the long run, we would like to replicate the results on larger portions of the resting state data, which would require a faster implementation. This should be possible using massively parallel architectures such as graphical processing units, as there are steps in the algorithm that are parallelizeable.

References

[1] R. M. Hutchison et al., "Dynamic functional connectivity: promise, issues, and interpretations," *Neuroimage*, vol. 80, pp. 360–378, 15 Oct. 2013.

[2] N. Leonardi and D. Van De Ville, "On spurious and real fluctuations of dynamic functional connectivity during rest," *Neuroimage*, vol. 104, pp. 430–436, 1 Jan. 2015.

[3] M. Breakspear, "Dynamic models of large-scale brain activity," *Nat. Neurosci.*, vol. 20, pp. 340–352, 23 Feb. 2017.

[4] J. Ou, L. Xie, C. Jin, X. Li, D. Zhu, R. Jiang, Y. Chen, J. Zhang, L. Li, and T. Liu, "Characterizing and differentiating brain state dynamics via hidden markov models," *Brain Topogr.*, vol. 28, pp. 666–679, Sept. 2015.

[5] Y. Du et al., "Identifying dynamic functional connectivity biomarkers using GIG-ICA: Application to schizophrenia, schizoaffective disorder, and psychotic bipolar disorder," *Hum. Brain Mapp.*, 10 Mar. 2017.

[6] E. A. Allen, E. Damaraju, S. M. Plis, E. B. Erhardt, T. Eichele, and V. D. Calhoun, "Tracking whole-brain connectivity dynamics in the resting state," *Cereb. Cortex*, vol. 24, pp. 663–676, Mar. 2014.

[7] U. Sakoğlu, G. D. Pearlson, K. A. Kiehl, Y. M. Wang, A. M. Michael, and V. D. Calhoun, "A method for evaluating dynamic functional network connectivity and task-modulation: application to schizophrenia," *MAGMA*, vol. 23, pp. 351–366, Dec. 2010.

[8] R. Hindriks, M. H. Adhikari, Y. Murayama, M. Ganzetti, D. Mantini, N. K. Logothetis, and G. Deco, "Can sliding-window correlations reveal dynamic functional connectivity in resting-state fMRI?," *Neuroimage*, vol. 127, pp. 242–256, 15 Feb. 2016.

[9] S. Shakil, C.-H. Lee, and S. D. Keilholz, "Evaluation of sliding window correlation performance for characterizing dynamic functional connectivity and brain states," *Neuroimage*, vol. 133, pp. 111–128, 4 Mar. 2016.

[10] T. O. Laumann et al., "On the stability of BOLD fMRI correlations," *Cereb. Cortex*, 2 Sept. 2016.

[11] D. Vidaurre, A. J. Quinn, A. P. Baker, D. Dupret, A. Tejero-Cantero, and M. W. Woolrich, "Spectrally resolved fast transient brain states in electrophysiological data," *Neuroimage*, vol. 126, pp. 81–95, 26 Nov. 2015.

[12] S. Ryali, K. Supekar, T. Chen, J. Kochalka, W. Cai, J. Nicholas, A. Padmanabhan, and V. Menon, "Temporal dynamics and developmental maturation of salience, default and Central-Executive network interactions revealed by variational bayes hidden markov modeling," *PLoS Comput. Biol.*, vol. 12, p. e1005138, Dec. 2016.

[13] S. F. V. Nielsen, K. H. Madsen, R. Røge, M. N. Schmidt, and M. Mørup, "Nonparametric modeling of dynamic functional connectivity in fMRI data," 4 Jan. 2016.

[14] S. Hidot and C. Saint-Jean, "An Expectation–Maximization algorithm for the wishart mixture model: Application to movement clustering," *Pattern Recognit. Lett.*, vol. 31, pp. 2318–2324, 15 Oct. 2010.

[15] J. Korzen, K. H. Madsen, and M. Mørup, "Quantifying temporal states in rs-fMRI data using bayesian nonparametrics." Poster presentation at Human Brain Mapping 2014, 2014.

[16] A. Cherian, V. Morellas, and N. Papanikolopoulos, "Bayesian nonparametric clustering for positive definite matrices," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, pp. 862–874, May 2016.

[17] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[18] R. A. Poldrack et al., "Long-term neural and physiological phenotyping of a single human," *Nat. Commun.*, vol. 6, p. 8885, 9 Dec. 2015.

[19] A. X. Patel, P. Kundu, M. Rubinov, P. S. Jones, P. E. Vértes, K. D. Ersche, J. Suckling, and E. T. Bullmore, "A wavelet method for modeling and despiking motion artifacts from resting-state fMRI time series," *Neuroimage*, vol. 95, pp. 287–304, 15 July 2014.

A.3 Predictive Assessment of Models for Dynamic Functional Connectivity

Nielsen, Søren F V, Mikkel N Schmidt, Kristoffer H Madsen, and Morten Mørup (2018). "Predictive assessment of models for dynamic functional connectivity". en. In: *Neuroimage* 171, pp. 116–134. ISSN: 1053-8119, 1095-9572. DOI: 10.1016/j.neuroimage.2017.12.084.

NeuroImage 171 (2018) 116–134

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Predictive assessment of models for dynamic functional connectivity

Søren F.V. Nielsen [a,*], Mikkel N. Schmidt [a], Kristoffer H. Madsen [a,b], Morten Mørup [a]

[a] DTU Compute, Technical University of Denmark, Richard Petersens Plads, Building 324, DK-2800 Kgs. Lyngby, Denmark
[b] Danish Research Centre for Magnetic Resonance, Section 714, Copenhagen University Hospital Hvidovre, Kettlegaard Allé 30, DK-2650 Hvidovre, Denmark

ARTICLE INFO

Keywords:
Dynamic functional connectivity
Hidden Markov models
Predictive likelihood

ABSTRACT

In neuroimaging, it has become evident that models of dynamic functional connectivity (dFC), which characterize how intrinsic brain organization changes over time, can provide a more detailed representation of brain function than traditional static analyses. Many dFC models in the literature represent functional brain networks as a meta-stable process with a discrete number of states; however, there is a lack of consensus on how to perform model selection and learn the number of states, as well as a lack of understanding of how different modeling assumptions influence the estimated state dynamics. To address these issues, we consider a predictive likelihood approach to model assessment, where models are evaluated based on their predictive performance on held-out test data. Examining several prominent models of dFC (in their probabilistic formulations) we demonstrate our framework on synthetic data, and apply it on two real-world examples: a face recognition EEG experiment and resting-state fMRI. Our results evidence that both EEG and fMRI are better characterized using dynamic modeling approaches than by their static counterparts, but we also demonstrate that one must be cautious when interpreting dFC because parameter settings and modeling assumptions, such as window lengths and emission models, can have a large impact on the estimated states and consequently on the interpretation of the brain dynamics.

Introduction

The functional integration of the brain can be studied by analyzing the patterns of synchronized activity across spatially separated brain regions. It has recently been shown that the functional connectivity (FC) varies with time, and a number of studies have investigated this dynamic functional connectivity (dFC) both in magneto/electro-encephalography (M/EEG) and functional magnetic resonance imaging (fMRI) (see recent reviews by Hutchison et al., 2013; Calhoun et al., 2014; Calhoun and Adali, 2016; O'Neill et al., 2017).

dFC can be studied by computing a static measure of FC (such as Pearson correlation) for successive windowed segments of activation time series. In accordance with the idea of meta-stability in the brain, cluster analysis (e.g. using the k-means algorithm) of the dFC time courses can then be used to identify a smaller set of FC *states* that occur repeatedly across time (Allen et al., 2014). A challenge with this windowed k-means (WKM) approach is that it is sensitive to the selection of the window length (Shakil et al., 2016; Hindriks et al., 2016) which implicitly defines the time scale of the dFC.

As an alternative to WKM, a window free approach based on a hidden Markov model (HMM) has recently been proposed (Baker et al., 2014; Ryali et al., 2016; Vidaurre et al., 2017a, 2017b; Nielsen et al., 2016). A HMM is a probabilistic sequence model which assigns a state label to each time point in the activation time series. The transitions between states are governed by a Markov process, and each state is characterized by a probability distribution over possible observations (which we refer to as the *emission model*). The state sequence, transition probabilities, and parameters of the emission model are estimated jointly when fitting the model. Analyzing resting state MEG power envelopes, Baker et al. (2014) proposed using a multivariate Gaussian emission model with state specific mean and covariance. A more advanced state-specific vector auto-regressive (VAR) emission model was considered by Vidaurre et al. (2016) to model raw MEG time series, in which each state also captures frequency structure and interactions in terms of a multivariate set of autoregressive coefficients. In resting state fMRI, the HMM with Gaussian emission model has been used in Ryali et al. (2016); Vidaurre et al. (2017b). The sliding window and HMM-based approaches to modeling dFC are illustrated in Fig. 1.

Several studies have investigated the statistical support for the assumption of dynamic changes in FC. Using an auto-regressive model of pairwise connections between brain nodes, Zalesky et al. (2014) found that relatively few connections were in fact dynamic but that there was support for dFC in resting state fMRI. Using a sinusoidal model, Leonardi & Van De Ville (2015) demonstrated how spurious fluctuations in FC

* Corresponding author.
E-mail addresses: sfvn@dtu.dk (S.F.V. Nielsen), mnsc@dtu.dk (M.N. Schmidt), kristofferm@drcmr.dk (K.H. Madsen), mmor@dtu.dk (M. Mørup).

https://doi.org/10.1016/j.neuroimage.2017.12.084
Received 8 September 2017; Accepted 24 December 2017
Available online 30 December 2017

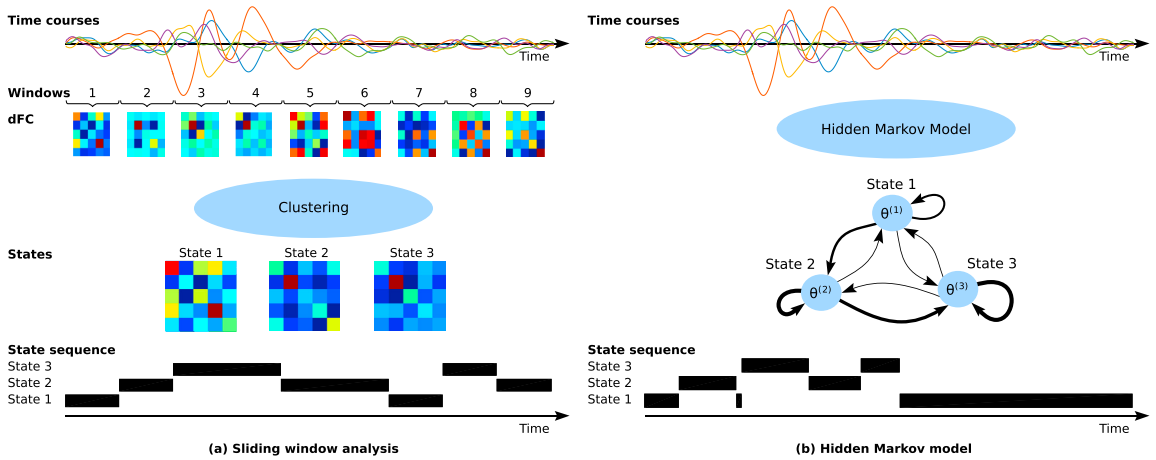


Fig. 1. Overview of the sliding window approach and hidden Markov model for extracting dFC. In this example both models were fitted on ERP-data from one subject (see section 3.3) using independent component analysis (ICA) time-courses of the neuronal signal from a number of brain regions. (a) In the sliding window approach, we divided the input time courses into 9 non-overlapping windows, each with 50 time points, and then computed the correlation matrix for each window. Next, we clustered the correlation matrices using k-means clustering with 3 components. (b) The hidden Markov model was fitted directly to the time courses using a multivariate Gaussian emission model and 3 states.

could arise due to model mismatch, and concluded that an appropriate window length was around 100 s based on the slowest frequency component of the BOLD signal. However, as Zalesky and Breakspear (2015) points out the sinusoidal model does not capture the correct spectral properties of fMRI data, and the conclusion is that more sophisticated generative models are needed. Laumann et al. (2016) conclude in their paper on stability of the BOLD signal that some of the dynamics observed can be attributed to head motion and subjects falling asleep in the scanner, but that some of the neural signal still remains unexplained.

While dFC analysis has become a widely accepted approach to analyze functional neuroimaging data, important open problems remain, including determining the number of brain states, and for sliding window methods to determine the window length. While a HMM can estimate the appropriate time scale from data, it is unclear how to best define the emission model. Since these modeling choices can significantly influence the interpretation of dFC, we posit there is a demand for a principled approach to compare different models of dFC.

In this paper we present a predictive model validation method in which dFC models are assessed based on their ability to characterize previously unseen data from the same experiment. To predict held-out data in a principled and quantifiable manner, we take a fully probabilistic modeling approach. While HMMs are probabilistic by nature, the WKM approach is not. We therefore consider WKM within a probabilistic setting by reformulating it as a Wishart mixture model (WMM) (Hidot and Saint-Jean, 2010; Korzen et al., 2014; Cherian et al., 2016; Nielsen et al., 2017). Within these probabilistic model specifications we use predictive validation to estimate the appropriate model complexity, including the appropriate number of brain states within each model specification. We thereby quantify whether or not functional connectivity is dynamic: "Does the data support more than one state?", as well as the complexity of dFC: "How many states best account for the held-out data?" in a data-driven way. For dFC specified by HMMs, we use our predictive assessment method to systematically investigate the influence of different emission models on the number of estimated states as well as on their ability to characterize held-out functional data.

We hypothesize that dynamics in dFC-models are strongly influenced by model parameters such as window lengths, emission models, and model order, and that the more complicated emission models will be able to explain the data at hand using fewer states compared to the simpler emission models. We demonstrate this using our predictive assessment framework on both synthetic dFC data with ground truth as well as real publicly available EEG (Wakeman and Henson, 2015) and fMRI data (Poldrack et al., 2015).

Methods

In the following, we examine four different models: a probabilistic formulation of the WKM as well as three hidden Markov models with different emission models. We treat all models in a non-parametric Bayesian setting (Orbanz and Teh, 2011): Using a prior distribution for the states based on a Dirichlet process (DP) allows us to estimate both the state parameters as well as the number of states simultaneously from data. We formulate the WKM approach as a DP-mixture model (Rasmussen, 1999) with Wishart distributed observations of windowed covariance matrices, and we consider three Gaussian DP-HMMs (Beal et al., 2002) with state-specific covariance and i) zero mean, ii) state-specific mean, and iii) state-specific vector auto-regressive mean. These non-parametric Bayesian models are commonly referred to as "infinite", as they can be derived by taking a limit as the number of states goes to infinity in a corresponding finite state model. Although these models support an unbounded number of states, inference on a finite data set will invoke only a finite subset, thus providing a statistically well founded mechanism for estimating the number of states. We further contrast this approach to the more conventional finite, parametric modeling approach as implemented by Vidaurre et al. (2016) (see also appendix section B).

The infinite wishart mixture model (IWMM)

The windowed k-means (WKM) approach has been used extensively in the dFC literature (Allen et al., 2014; Rashid et al., 2016). Small "snapshots" of connectivity are estimated using L sliding windows and the snapshots are represented as correlation matrices, $\mathbf{\Omega}_\ell$, for each window ℓ. To find common connectivity patterns the upper triangular part of each $\mathbf{\Omega}_\ell$ is stacked into a vector, $\boldsymbol{\omega}_\ell$, and finally k-means clustering is performed on the collection of vectors $\{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_L\}$ using K clusters and the Euclidean distance measure. A common problem in the WKM is how to choose K, and heuristics such as the elbow-criterion are often used.

To be able to perform predictive validation on previously unseen data, and to learn the number of clusters as part of the model inference,

we reformulate the WKM approach as a probabilistic generative model. Each windowed covariance matrix $\mathbf{\Omega}_\ell$ is the mean-subtracted scatter matrix, $\boldsymbol{C}_\ell$, of the data within each window segment ℓ, defined as

$$\boldsymbol{C}_\ell = \sum_t w_\ell(t)\mathbf{x}_t\mathbf{x}_t^T, \tag{1}$$

where $\mathbf{x}_t \in \mathbb{R}^p$ is the data vector at time t and $\boldsymbol{w}_\ell(t)$ is the window function associated with the ℓth window. Under a multivariate Gaussian assumption and rectangular windows, the scatter matrices follow a Wishart distribution, and a clustering of these can be achieved using a Wishart mixture model (WMM) as proposed by Hidot and Saint-Jean (2010). We argue that the WMM is the most natural and direct probabilistic formulation of the WKM approach. We presently consider the DP-mixture version of the WMM, the so-called infinite Wishart mixture model (IWMM), as proposed by Korzen et al. (2014).

The IWMM assumes that each state has an associated covariance matrix $\mathbf{\Sigma}_k$, drawn from an inverse Wishart prior, and that each observed data window belongs to one of the K states, where K lies between one and the number of observations. In the DP-mixture, the prior distribution over the state assignments is given by the so-called Chinese restaurant process (CRP) (Aldous, 1985); a distribution that has support on all state assignments corresponding to all possible partitions of the observations. This yields the following generative model for the IWMM,

$$\mathbf{z} \sim \mathrm{CRP}(\alpha), \tag{2}$$

$$\mathbf{\Sigma}_k \sim \mathscr{W}^{-1}(\mathbf{\Sigma}_0, \nu_0), \tag{3}$$

$$\mathbf{C}_\ell \sim \mathscr{W}\left(\mathbf{\Sigma}_{z_\ell}, \nu_\ell\right), \tag{4}$$

in which $\mathbf{z}$ is the state assignment of each window, $\mathbf{\Sigma}_0$ is the prior covariance with ν_0 degrees of freedom and ν_ℓ is the degrees of freedom for the ℓth windowed covariance matrix (in the case of a rectangular window this is equal to the window length). Due to conjugacy between the Wishart and inverse Wishart distribution we can marginalize out all the $\mathbf{\Sigma}_k$'s and carry out the inference in terms of the state assignment parameters only. In the IWMM we parameterize the prior $\mathbf{\Sigma}_0 = \eta\mathbf{I}$, in which η is a positive scaling parameter that determines the strength of the prior.

One problem still persist for WKM and IWMM, namely how to choose the window-length. We cannot compare models using predictive likelihood across different window-lengths since the likelihood function itself depends on the window length. The most flexible choice of window length is 1, in which we arrive at a likelihood function proportional to a Gaussian mixture model (GMM), but here we lose much of the stability and robustness achieved with longer window lengths. To model a slowly changing state sequence, the most natural extension is thus to use a hidden Markov model (HMM), which we discuss in the following.

The infinite hidden Markov model

In neuroimaging, hidden Markov models have frequently been used for modeling dFC (Baker et al., 2014; Vidaurre et al., 2016, 2017a; Ryali et al., 2016; Nielsen et al., 2016,b). In a manner similar to a DP-mixture model, the non-parametric version of the HMM, termed the infinite HMM (IHMM) (Beal et al., 2002), learns the number of states as part of the inference. The generative model for the IHMM can be written as,

$$b_k \sim \mathrm{Beta}(1, \gamma), \tag{5}$$

$$\beta_k = b_k\prod_{\ell=1}^{k-1}(1-b_\ell), \tag{6}$$

$$\boldsymbol{\pi}^{(k)}\big|\boldsymbol{\beta} \sim \mathrm{DP}(\alpha, \boldsymbol{\beta}), \tag{7}$$

$$z_t\big|z_{t-1} \sim \mathrm{Multinomial}\left(\boldsymbol{\pi}^{(z_{t-1})}\right), \tag{8}$$

$$\theta^{(k)} \sim H, \tag{9}$$

$$\mathbf{x}_t \sim F\left(\theta^{(z_t)}\right), \tag{10}$$

in which γ and α are positive parameters, $\boldsymbol{\beta}$ is a vector of infinite length (in practice one needs only to work with a finite representation), $\boldsymbol{\pi}$ is the transition matrix with rows $\boldsymbol{\pi}^{(k)}$ and DP is the Dirichlet process (Ferguson, 1973) — a distribution over discrete probability distributions, parameterized by a base measure $\boldsymbol{\beta}$ and a concentration parameter α (for a thorough exposition of the DP, see e.g. Blei and Jordan, 2006; Van Gael, 2011). The state specific parameters, $\theta^{(k)}$, are assumed to be drawn from a here unspecified prior distribution H, and data is drawn from the unspecified distribution F dependent on which state that particular data point, $\boldsymbol{x}_t$, belongs to. A graphical model for the IHMM can be seen in Figure S.1b in the appendix.

Emission models

We investigate three emission models of increasing complexity that have previously been used for modeling dFC: a zero-mean Gaussian (ZMG) (Nielsen et al., 2016), a Gaussian with a state-specific mean (SSM) (Rezek and Roberts, 2005; Baker et al., 2014), and Gaussian with an auto-regressive mean (VAR) (Fox et al., 2011; Vidaurre et al., 2016). In all cases the covariance is state-specific and models that state's functional connectivity. There are other emission models in the Gaussian family such as the state specific mean model with isotropic variance (Baldassano et al., 2017) and other variants where the covariance is constrained. These will not be considered presently because they do not model the full functional connectivity. The emission parameters are distributed as described in Table 1.

For all the HMM emission models we have chosen conjugate distributions, to be able to analytically marginalize $\mathbf{\Sigma}^{(k)}$, $\boldsymbol{\mu}^{(z_t)}$, and $\mathbf{A}^{(z_t)}$, such that inference is carried out on the state sequence alone.

Predictive likelihood

To assess and compare the different models, we examine their ability to generalize, i.e., how well a model fitted on training data, $\mathbf{X}$, can account for unseen test data, $\mathbf{X}^*$, from the same experiment or paradigm. This can be viewed as an alternative to classical statistical inference and hypothesis testing (Bzdok and Yeo, 2017).

Thus we are interested in evaluating the following integral,

$$p(\mathbf{X}^*|\mathbf{X}, \mathscr{M}) = \int_{\mathbf{\Theta}\in\mathscr{M}} p(\mathbf{X}^*|\mathbf{\Theta})p(\mathbf{\Theta}|\mathbf{X}), \tag{11}$$

yielding the *posterior predictive likelihood* (from now on denoted the predictive likelihood), in which $\mathbf{\Theta} \in \mathscr{M}$ is the collection of all model parameters and $\mathscr{M}$ is the model space. The predictive likelihood

Table 1
Emission models used in the HMMs where $\mathbf{\Sigma}^{(k)}$ is the state-specific $p \times p$ covariance matrix, $\mathscr{W}^{-1}$ is the inverse Wishart distribution, Σ_0 is the prior covariance, ν_0 is the degrees of freedom (in all experiments $\nu_0 = p$), $\boldsymbol{\mu}_0$ is the prior mean of the signal, λ is a positive precision parameter of the mean, $\mathscr{MN}(M, U, V)$ is the matrix-normal distribution with mean M, row-variance U and column variance V, $\mathbf{A}^{(k)}$ is a $p \times pr$ matrix containing the coefficients for the k'th state of an order r VAR process, and $\bar{\mathbf{x}}_t$ are the r-lagged observations for time point t stacked in a vector.

Zero Mean Gaussian	State-Specific Mean	Vector Autoregressive
ZMG	SSM	VAR
$\mathbf{\Sigma}^{(k)} \sim \mathscr{W}^{-1}(\mathbf{\Sigma}_0, \nu_0)$	$\mathbf{\Sigma}^{(k)} \sim \mathscr{W}^{-1}(\mathbf{\Sigma}_0, \nu_0)$	$\mathbf{\Sigma}^{(k)} \sim \mathscr{W}^{-1}(\mathbf{\Sigma}_0, \nu_0)$
	$\boldsymbol{\mu}^{(k)} \sim \mathscr{N}(\boldsymbol{\mu}_0, \lambda^{-1}\mathbf{\Sigma}^{(k)})$	$\mathbf{A}^{(k)} \sim \mathscr{MN}(\mathbf{0}, \mathbf{\Sigma}^{(k)}, \mathbf{I})$
$\mathbf{x}_t \sim \mathscr{N}(\mathbf{0}, \mathbf{\Sigma}^{(z_t)})$	$\mathbf{x}_t \sim \mathscr{N}(\boldsymbol{\mu}^{(z_t)}, \mathbf{\Sigma}^{(z_t)})$	$\mathbf{x}_t \sim \mathscr{N}(\mathbf{A}^{(z_t)}\bar{\mathbf{x}}_t, \mathbf{\Sigma}^{(z_t)})$

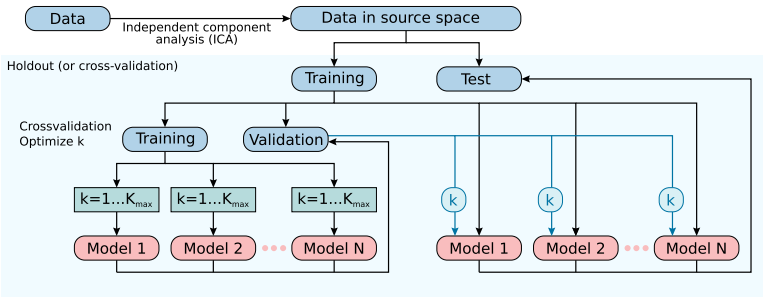


Fig. 2. A schematic overview of the predictive likelihood framework, that visualizes the nested cross-validation framework. In this figure the models of dynamic functional connectivity (dFC) can be anything, as long as the predictive likelihood on held-out data can be estimated. If the likelihood function is the same across models we can use this framework to do model selection.

quantifies the probability of observing the test data under the model given the training data and the model space, and can be viewed as the likelihood of the test data averaged over the posterior distribution of the parameters fitted on the training data. To showcase that the predictive likelihood framework is also applicable for other probabilistic models we also use the (finite) variational Bayesian HMM (VB-HMM) from Vidaurre et al. (2016).[1] A description of the model can be seen in appendix section B, along with details on how to calculate the predictive likelihood for all models.

To use predictive evaluation, the data must be divided into independent training and test sets. In dFC, where the data is modeled as sequential, this can be done by splitting the time series into subsequences. Alternatively, it is possible to train the model on whole time series from one or more subjects, and use data from independent, held-out subjects for testing. In this paper we use the predictive likelihood to do model selection and parameter tuning in a two level cross-validation framework. In the inner part, we estimate the prior strength η for the IWMM and IHMMs considered, and the number of states for VB-HMM for all emission models. In the outer part, we quantify each of the emission model's capability of explaining the held-out test data. We emphasize that we cannot directly compare the predictive likelihood across IWMM, VB-HMM and IHMM. The IWMM uses a different likelihood function than the two HMM-models, i.e. the IWMM models covariance matrices as the observed quantity whereas the HMMs model the time series directly. For the VB-HMM we have chosen to use a VB-bound to approximate the integral in (11) (Beal, 2003), that has the advantage of propagating the uncertainty in the parameters from training at the cost of estimating the state-sequence distribution on the test set. In the IHMM we use samples from the posterior obtained during training together with Viterbi integration (more details on this can be found in Appendices B, C and D). A general schematic of the predictive likelihood framework can be seen in Fig. 2.

Of present interest is to investigate under a given independent component analysis (ICA) representation which model of dFC most adequately describes this representation. We therefore consider the ICA as a preprocessing step applied to all the data. Alternatively, the ICA could have been applied separately on the training and test data. Training the ICA independently on the training and test set would result in an issue of matching components (Du et al., 2012), whereas defining the ICA only on the training data and projecting the test data onto these learned components can result in issues of variance inflation (Abrahamsen and Hansen, 2011). By considering the ICA as a preprocessing step we remove any influence that changes in the ICA representation across training and test data may have. We are thereby not affected by these potential confounds and are able to quantify within a given ICA representation which model of dFC best characterizes the data.

For the remainder of this paper we will contrast the predictive likelihood of a model of interest versus a baseline model using the Bayes factor (Kass and Raftery, 1995; Nielsen et al., 2017), denoted *BF*. This can be calculated as,

$$\mathrm{BF} = \frac{p(\mathbf{X}^*|\mathbf{X}, \mathscr{M})}{p(\mathbf{X}^*|\mathbf{X}, \mathscr{M}_0)}, \tag{12}$$

in which $\mathscr{M}$ is the model of interest and $\mathscr{M}_0$ is the baseline model. Typically the baseline model will be the model with only one state, and thus the Bayes factor quantifies how much better it is to use a particular dynamic model. The Bayes factor is often used in the dynamic causal modeling (DCM) framework (Penny et al., 2004) to do model selection, however, an important distinction between the DCM and our approach is that the BF in DCM in calculated on the training data whereas the BF in this paper is calculated on held-out test data.

Evaluating similarity of state sequences

To compare different models, we also examine how similar their estimated state sequences are. Here, we use normalized mutual information (NMI) to quantify the correspondence of two different sequences, possibly with differing number of states. Let the state sequences be given by state assignments vectors $\boldsymbol{z}^{(a)}$ and $\boldsymbol{z}^{(b)}$. Then, the NMI is given by

$$\mathrm{NMI}\left(\boldsymbol{z}^{(a)}, \boldsymbol{z}^{(b)}\right) = \frac{2\mathrm{MI}\left(\boldsymbol{z}^{(a)}, \boldsymbol{z}^{(b)}\right)}{\mathrm{MI}\left(\boldsymbol{z}^{(a)}, \boldsymbol{z}^{(a)}\right) + \mathrm{MI}\left(\boldsymbol{z}^{(b)}, \boldsymbol{z}^{(b)}\right)}, \tag{13}$$

where MI is the mutual information.

Experiments and results

The proposed approach for predictive assessment of dFC models was validated on synthetic data, and demonstrated on two real data sets based on electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) as described in the following sections.

The influence of window lengths

A challenge in the WKM approach as well as its probabilistic formulation, the IWMM, is the specification of a suitable window length (Shakil et al., 2016; Zalesky and Breakspear, 2015; Leonardi & Van De Ville, 2015; Hindriks et al., 2016). If the window length is too short, the windowed data will be less statistically stable and the approach might find spurious states. If, on the other hand, the window length is too long, short-lived states might not be detectable. In contrast, the HMM approach does not involve windowed analysis.

[1] MATLAB code was downloaded from the repository https://github.com/OHBA-analysis/HMM-MAR in July 2016. The predictive likelihood code was written by the authors.

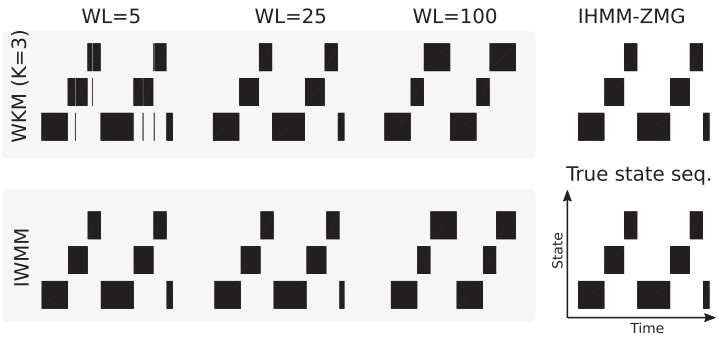


Fig. 3. Estimated and true state sequences for synthetic data I experiment. Data were generated from a three-state model, where each states had a differing covariance matrix. Results are shown for windowed k-means (WKM) and infinite Wishart mixture model (IWMM) with window lengths of 5, 25, and 100 samples as well as the infinite hidden Markov model with zero mean Gaussian emission model (IHMM-ZMG).[1].

We applied WKM as well as the IWMM and IHMM to synthetic data with ground truth in order to investigate the merits of windowed covariance modeling versus HMMs that do not assume a priori time-windowing but learns the state dynamics and their smoothness as part of the inference. In all analysis the AR-order was set to 1 in both the IHMM and the VB-HMM (see more about this choice in the Discussion).

Synthetic data I. We generated two data sets (training and validation) from the same 5-dimensional 3-state ZMG model, i.e., the data were generated to have three different states, defined by different covariance matrices, in a fixed state sequence. The covariance matrices for each state were generated as $\mathbf{U}\mathbf{U}^{\top}$ where $\mathbf{U}$ was an upper triangular matrix with i.i.d. standard Gaussian entries. In each data set, the total length of the generated time series was 500 samples (i.e., if this was fMRI we would have had 500 TRs), and the state sequence was chosen such that the states had different durations with the shortest state occurrence lasting 50 samples.

Synthetic experiment I. For WKM we set the number of states to the true number of states ($K = 3$). For IWMM and IHMM we tuned the prior covariance scale parameter η by fitting the models on the training data and optimizing the parameter using predictive likelihood on the validation data. We then concatenated the training and validation data, and using the full data set with 1000 time points we fitted the WKM, IWMM, and IHMM-ZMG models. For the WKM and the IWMM we used rectangular non-overlapping windows, and compared window lengths of 5, 25, and 100 samples chosen to represent a too short, an appropriate (i.e. one that does not mix together different states), and a too long window.

Results on synthetic data I. The results can be seen in Fig. 3 which shows the estimated state sequences. The WKM and IWMM perform almost identically: When the window length is appropriate (WL = 25) both methods detect the correct state sequence. When the window length is too large (WL = 100) both fail to capture the short-lived state correctly, and when it is too small (WL = 5) the WKM detects spurious states. Both IWMM and the IHMM-ZMG correctly identify the number of states using the cross-validated value of η. Furthermore, the IHMM captures the true state sequence without a priori specifying and averaging across windows. Thus, all models can correctly identify the underlying dFC on data in compliance with their assumptions. It should be noted that adequately tuned *overlapping* and *tapered windows* (Allen et al., 2014) could potentially alleviate the issues encountered using too long window lengths, however, this was not considered in this experiment.

HMM emission models

In the hidden Markov model approach to estimating dFC, we claim that the choice of emission model can have a large influence on the result. To substantiate this, we compared the three examined emission models by performing a pair-wise comparison investigating how well each model was able to estimate the true state sequence on synthetic data generated according to each of the three model specifications. Furthermore, we compared how well each model was able to characterize dFC by computing the predictive likelihood on held-out validation data.

Synthetic data II. We generated synthetic data from each of the three emission models (ZMG, SSM, and VAR) with five dimensions and three states (we used the same state sequence as in the previous synthetic experiment shown in Fig. 3). Training, validation, and test data sets were generated with identical parameter settings for each data model. For all models, the covariance matrix for each state was defined as in the previous synthetic experiment. For the SSM model, the state-specific means (5-dimensional vectors) were generated randomly with i.i.d. standard Gaussian entries. The state-specific VAR coefficients were generated, by first generating a p-dimensional signal from a sinusoid with random frequency (common for all dimensions) and random phase (different for each dimension). We then fitted a VAR-model of order 1 to that (using the least squares estimator) and finally generated new data from the fitted model with i.i.d. standard Gaussian noise.

Synthetic data experiment II. For IWMM and IHMM, the prior strength η was selected by cross-validation using the training and validation set, and the models were then fitted on the concatenated training and validation data. The predictive likelihood was computed for each of the fitted models using the test data. For comparison we also fitted the WKM model, both with the correct number of clusters ($K = 3$) and with too many clusters ($K = 6$). Both the WKM's and IWMM were run with an appropriate window length of 25. To investigate the influence of the inference procedure, we also fitted the models using the VB-HMM implementation by Vidaurre et al. (2016).

Results on synthetic data II. The estimated state sequences for each of the fitted models are shown in Fig. 4. When the number of states was specified correctly ($K = 3$) the WKM found the true state sequence for all three data sets; however, when the number of states was misspecified ($K = 6$) the WKM failed in all cases and appeared to subdivide each state. The IWMM was able to learn the true state sequence for the ZMG and SSM-emission data, but failed in the case of the VAR-emission data. The three IHMM models found the true state sequence in the cases when the data were generated from one of the two simple emission models (ZMG and SSM), except the IHMM-VAR which falsely detected two single-time-point clusters for the SSM-data. When the data were generated from the VAR model, only the VAR model and the WKM with the correct number of clusters found the correct state sequence. In this setting, the IHMM-ZMG and IHMM-SSM both failed in estimating the true number of underlying states and detected multiple spurious states. This indicates that these more simple models needed more states (and parameters) to account for the more complex VAR data. Results for

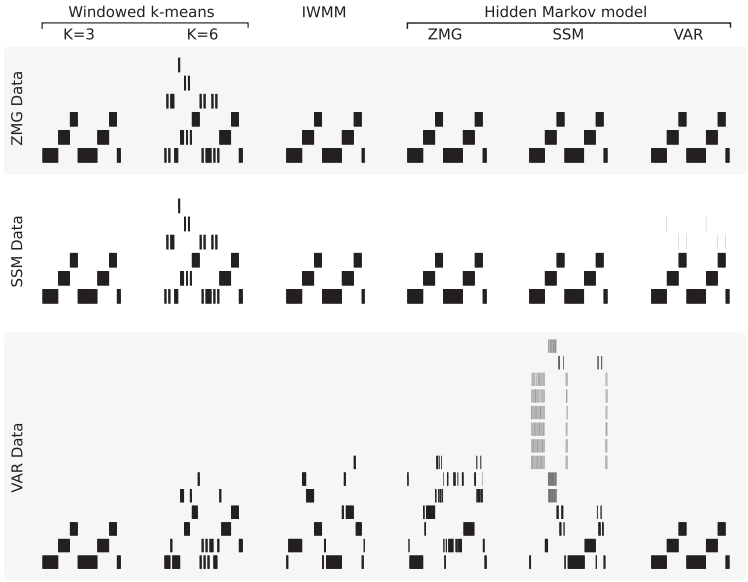


Fig. 4. Estimated state sequences for synthetic data II generated from hidden Markov models. Top: Zero mean Gaussian (ZMG) emission. Middle: State-specific mean (SSM) emissions. Bottom: Vector autoregressive (VAR) emission. Results are shown for windowed k-means (WKM) with $K = 3$ and $K = 6$ clusters, the infinite Wishart mixture model (IWMM), and infinite hidden Markov models with ZMG, SSM, and VAR emission models. The true state sequence is shown in Fig. 3.

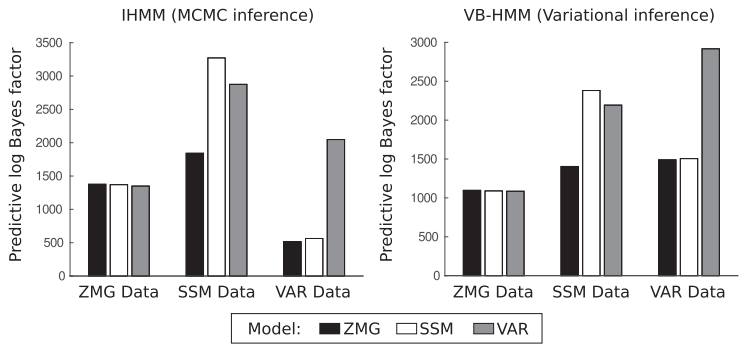


Fig. 5. Bayes factors (synthetic data) for each model vs. a baseline model containing only one zero-mean state (static functional connectivity) computed on held-out validation data. This was one using both Markov chain Monte Carlo inference (left) and variational Bayesian inference (right).

VB-HMM were similar to the IHMM and can be found in appendix section E.

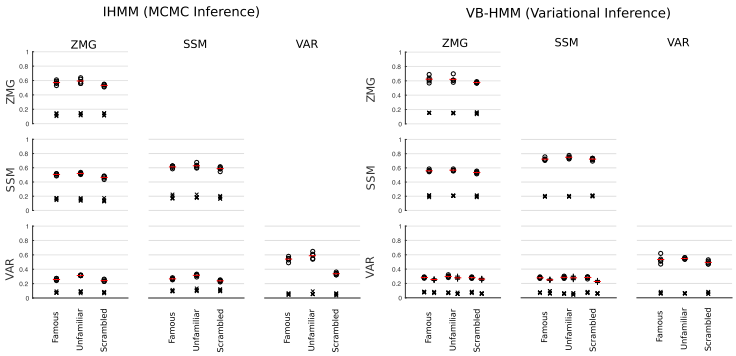
The predictive likelihood of each model is reported in Fig. 5, which shows the predictive Bayes Factor of each emission model vs. a baseline model given by a one state (non-dynamic) zero mean Gaussian defined by the empirical covariance matrix of the concatenated training and validation set. As expected when the HMM emission model matched the emission model of the generated data, the best Bayes factor was achieved. When the data were simple (from ZMG) the three emission-models performed approximately equal (with the ZMG performing best), indicating that the more complex models could adapt to the simple data but not vice versa. We also conducted an experiment to investigate the influence of noise and fMRI signal properties on the predictive results. This can be seen in appendix section G.

EEG task paradigm analysis

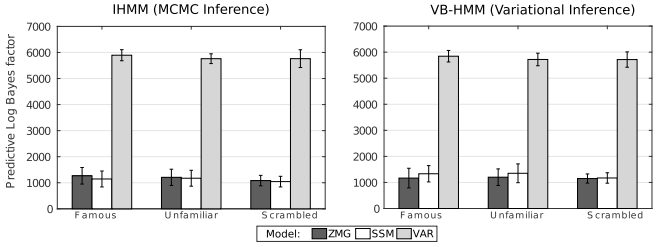
To verify that the proposed predictive evaluation framework produces sensible results, we demonstrate it on an electroencephalography (EEG) task-paradigm with very high signal-to-noise ratio using event related potentials (ERP), similar to the analysis carried out in Murray et al. (2008); Ott et al. (2011) under the name of topographical ERP mapping.

EEG data. We analyzed a publicly available face recognition task data set (Wakeman and Henson, 2015)[2] that consists of 16 subjects. The

[2] This data was obtained from the OpenfMRI database. Its accession number is ds000117. The preprocessing scripts for SPM were downloaded from ftp://ftp.mrc-cbu.cam.ac.uk/personal/rik.henson/wakemandg_hensonrn/.



(a) Normalized mutual information (NMI) between estimated state sequences (circles) for each data set and each pair of models. Within each model, the NMI measures the consistency of the estimated state sequences across five reruns of the inference algorithm. Between each pair of models, the NMI measures the similarity between the estimated state sequences. NMI computed against a random state sequence from the fitted model is shown as a baseline (crosses). Results are shown for inference using MCMC (left) and variational Bayes (right). For the variational Bayes (VB) models, the VAR was also compared to the ZMG and SSM model run with the same number of states as the VAR indicated by plusses in the third row of the VB-plot.



(b) Bayes factors (EEG data, *famous* condition) on test data from five held-out subjects against a baseline model containing only one zero-mean state (static functional connectivity).

Fig. 6. Comparison of model performance on the EEG-data. We plot the IHMM and VB-HMM performance in terms of model consistency and predictive likelihood on held-out data. (a) Normalized mutual information (NMI) between estimated state sequences (circles) for each data set and each pair of models. Within each model, the NMI measures the consistency of the estimated state sequences across five reruns of the inference algorithm. Between each pair of models, the NMI measures the similarity between the estimated state sequences. NMI computed against a random state sequence from the fitted model is shown as a baseline (crosses). Results are shown for inference using MCMC (left) and variational Bayes (right). For the variational Bayes (VB) models, the VAR was also compared to the ZMG and SSM model run with the same number of states as the VAR indicated by plusses in the third row of the VB-plot. (b) Bayes factors (EEG data, famous condition) on test data from five held-out subjects against a baseline model containing only one zero-mean state (static functional connectivity).

paradigm has three conditions: Either i) a *famous* face is presented, ii) an *unfamiliar* face is presented, or iii) a *scrambled* face (with the phase of the 2D-Fourier coefficients permuted) is presented. Our analysis was not focused on contrasting the conditions, and each condition was thus analyzed individually to investigate the robustness of the estimated dynamics.

The standard preprocessing, as described by Wakeman and Henson (2015) (which included low-pass filtering to 32 Hz), using the SPM8 MATLAB toolbox[3] was applied to the data and additionally we interpolated the automatically detected bad channels using the distance function in FieldTrip .[4] We then calculated individual event-related potentials (ERP) for each subject and condition and ran independent component analysis (ICA) on the concatenated data (all subjects and conditions) using the Infomax algorithm (Bell and Sejnowski, 1995) with five components. The number of components was chosen based on the eigenvalue spectrum of random uncorrelated data (Horn, 1965). An example of an ICA time course is displayed in lower right corner of Fig. 7.

EEG experiment. Eleven subjects were taken out for training, leaving five subjects for testing. In the training set five-fold cross-validation was applied to estimate the prior strength in the IHMM and the number of states for VB-HMM for each condition and each emission model using predictive log-likelihood on the validation set as a measure of fit. Each subject's ICA time courses from event related potentials (ERP) were concatenated in time, and to account for discontinuities in the data we set up the models to restart the state sequence at each new subject. After cross-validation, we re-trained the models on the whole training data and calculated the predictive likelihood on the test data.

To assess the robustness of the approach, we computed the normalized mutual information (NMI) of the estimated state sequences over five restarts of each model in the following manner: Restart (1 vs. 2), (2 vs. 3), (3 vs. 4), (4 vs. 5), and (5 vs. 1). To examine the similarity between the estimated state sequences across the three models, we computed the NMI between the models: Restart (1 vs. 1), (2 vs. 2) etc. for each pair of models (ZMG vs. SSM), (ZMG vs. VAR), and (SSM vs. VAR). As a baseline, each case was also compared to a null-model, in which one of the state sequences in each pair was replaced with a new state sequence sampled using the fitted transition matrix thus resulting in similar state transition dynamics as the original sequence but uninformed by the data.

EEG results. NMI scores comparing the estimated state sequences are given in Fig. 6a. Results for the three data sets (*familiar*, *unfamiliar*, and *scrambled*) were generally in close agreement with each other. For all

[3] http://www.fil.ion.ucl.ac.uk/spm/software/spm8/.
[4] http://www.fieldtriptoolbox.org/.

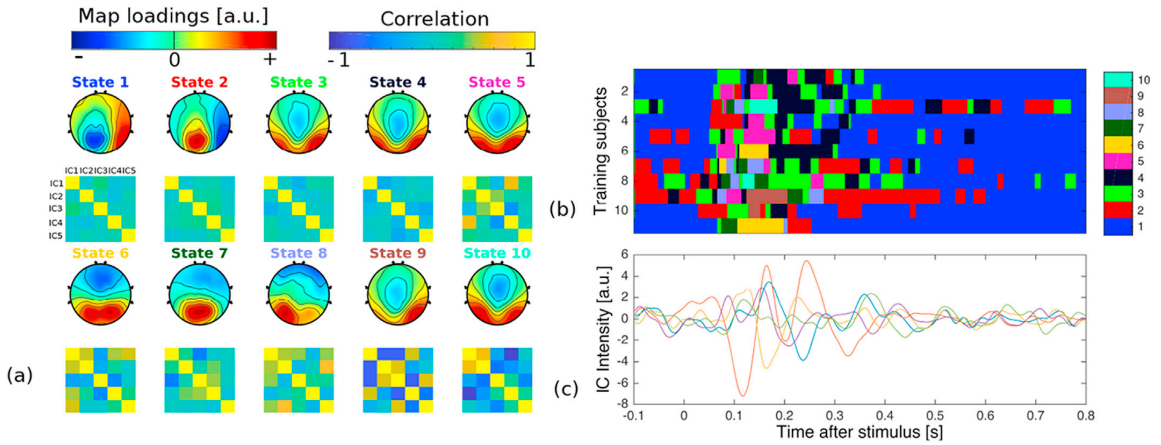


Fig. 7. Visualization of the best IHMM model, in this case the IHMM-VAR, according to the predictive framework for the "Famous" condition. (a) For each state we computed the first principal component of all the data points in the training set that belonged to that state and plotted that as a topographical map. Note that the states were ordered according to their fractional occupancy (largest state first). Below each map we plot the empirical correlation matrix of all data points assigned to the given state. (b) We plot for each timepoint the state-assignment for each subject as an image (each row represents a subject). Each color represents a state. (c) An example of one subjects data in ICA-space (each color represents a independent component).

models, NMI scores between restarts were higher than the baseline; thus, the estimated state sequences were relatively consistent over restarts, although all NMI scores were well below one, indicating some disagreement. NMI scores between ZMG and SSM were similar to NMI scores between restarts of the two models, indicating that the ZMG and SSM models estimated similar state sequences. NMI scores between VAR and the other two models were lower than NMI between restarts, indicating that the estimated state sequences for the VAR model were different from those estimated by the ZMG and SSM models. This was confirmed for the VB-HMM when running the ZMG and SSM models with the same number of states as the VAR model was run with, i.e., the state sequence obtained from the VAR model differs from the ZMG and SSM state sequences even if the ZMG and SSM have the same number of states as the VAR model. We also looked into the number of states estimated by each emission model; the VAR model estimated fewer states as hypothesized in the introduction with more smooth trajectories compared to the two other emission models (see appendix section J).

To investigate which emission model best characterized the held out subjects, the Bayes factor towards a baseline model (empirical covariance matrix of the training data) was calculated and can be seen in Fig. 6b. All models gave better performance than the baseline. The VAR emission model consistently gave best predictive performance across all conditions and for both inference methods.

For the best performing model, the IHMM-VAR, we take the best sample in terms of joint log-likelihood from the training inference, and visualize the solution in Fig. 7. To plot the topography of each state, we gathered all the time points assigned to a particular state and calculated the first principal direction, and plotted those values using EEGLAB. We did this to not be influenced by changes in polarity and because it resembles the microstate-analysis done in Khanna et al. (2015). We notice that there seems to be a baseline state (state 1), that some of the training subjects visit before stimulus and around 0.4 s after stimulus. In the period after stimulus (from 0.1 to 0.4 s) the dominant states' topography show high activity in the posterior areas consistent with the visual task. There seems to be a "consensus" of fewer states in the baseline (pre-stimulus and after 0.4 s after stimulus) and a larger number of different states being used right after stimulus. This indicates that we need more states to explain the difference in visual processing of faces across subjects compared to the baseline state. Furthermore, some states seem to have very similar topographical characteristics (i.e. states 3–5) but are different in their functional connectivity.

fMRI resting state analysis

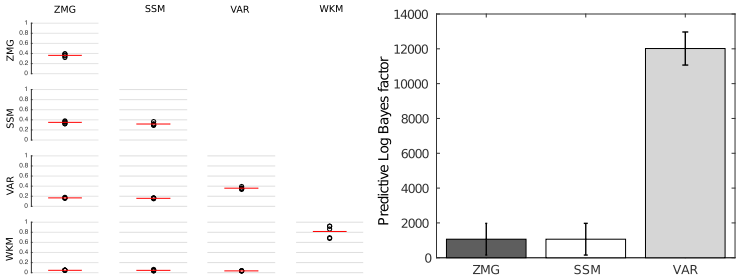
Finally, we will demonstrate our approach to predictive assessment of dFC models on a resting state fMRI data set. Subject variability can be a significant issue in dFC (Nielsen et al., 2016) and in neuroimaging in general (Finn et al., 2015) and care must be taken when interpreting dynamics at a group level, so we analyzed resting state fMRI data from a single subject. We contrast the extracted brain states from the HMM framework to those from sliding window k-means.

fMRI data. We used the resting state fMRI data from Poldrack et al. (2015)[5] which contains 89 recorded resting state fMRI sessions of a single subject. We applied the following pre-processing steps using SPM12[6]: We coregistered all sessions to the first image of the first functional session (session 014), and then jointly corrected all sessions for motion artifacts using a rigid-body transformation towards the mean volume. An anatomical image (T1W) from session 012 was used to segment grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF) using the standard tissue probability map from SPM. We applied a discrete cosine transform based bandpass filter with cut-off at $[0.009, 0.08]$ Hz to all sessions (as suggested in the methods section of (Poldrack et al., 2015)), along with nuisance regression of the motion parameters and mean signal within CSF and WM masks eroded by a 4 mm isotropic spherical kernel. We subsequently applied wavelet despiking (Patel et al., 2014) with standard parameters, and finally we resliced all sessions (due to a change in the number of slices after session 027) to the first session and smoothed using an isotropic 5 mm full width at half maximum Gaussian kernel. After preprocessing we ran a group ICA (Calhoun et al., 2001) implemented in the GIFT toolbox[7], using the ERBM algorithm with 30 components and otherwise default settings. We used 30 components, which can seem 'low' compared to other dFC analysees (Allen et al., 2014). However, this was done both for computational reasons, i.e. the HMM scales cubicly in the number components (cf. appendix B) and also for statistical reasons since we need enough degrees of freedom to

[5] This data was obtained from the OpenfMRI database. Its accession number is ds000031.

[6] http://www.fil.ion.ucl.ac.uk/spm/software/spm12/.

[7] http://mialab.mrn.org/software/gift/index.html.



(a) Normalized mutual information (NMI) between estimated state sequences on the resting-state fMRI training data both across restarts within each model and between the models (circles).

(b) Bayes factors (fMRI resting state data) computed on held-out test data for each model against a VB-HMM-ZMG with one state. Error-bars indicate standard deviation over test sessions.

Fig. 8. Comparison of model performance on the fMRI-data. We plot the VB-HMM performance in terms of model consistency and predictive likelihood on held-out data. For the model consistency in 8a we compare the HMMs to windowed k-means (WKM).

reliably estimate the covariance matrix of each state. We discarded 9 components based on visual inspection of the component spatial maps overlap with the brainstem and movement related effects, and thus ran the final HMM-analysis on the 21 remaining components. The retained components' spatial maps can be seen in appendix section K.

fMRI experiment. The data were only analyzed using the VB-HMM inference procedure due to the higher computational complexity of the IHMM. We split the 89 sessions randomly into two parts: 45 sessions for training and 44 sessions for testing. In the training set we performed five fold cross-validation to determine the number of states for all three emission models using the proposed predictive log-likelihood as a measure of fit. The final models were retrained five times on the training data, the best restart chosen by the minimum free-energy, and finally compared with predictive log-likelihood on the test sessions. To compare the estimated state sequences and assess the robustness of the approach, we conducted a NMI analysis as in the EEG experiment.

fMRI results. NMI scores comparing the estimated state sequences are shown in Fig. 8a. The NMI between state sequences estimated by the ZMG and SSM models were similar to NMI scores for restarts of the two models, indicating that the estimated state sequences were in agreement. NMI scores between VAR and the other two HMMs were lower, indicating that the VAR model found a different state sequence. We looked into the number of states estimated by the three emission models (see appendix section I) and found that the VAR identified six states, whereas the ZMG and SSM used 7 and 8 states respectively. From Fig. 8a it seems that the WKM found more robust results over restarts and was in very low agreement with the HMMs.

The predictive performance on the test set for each of the models is given in Fig. 8b, which shows log Bayes factors against a baseline given by the ZMG one state model. As in the EEG analysis the VAR model outperformed the other models in terms of predictive likelihood.

Finally, we visualize the states from best performing model (i.e. the

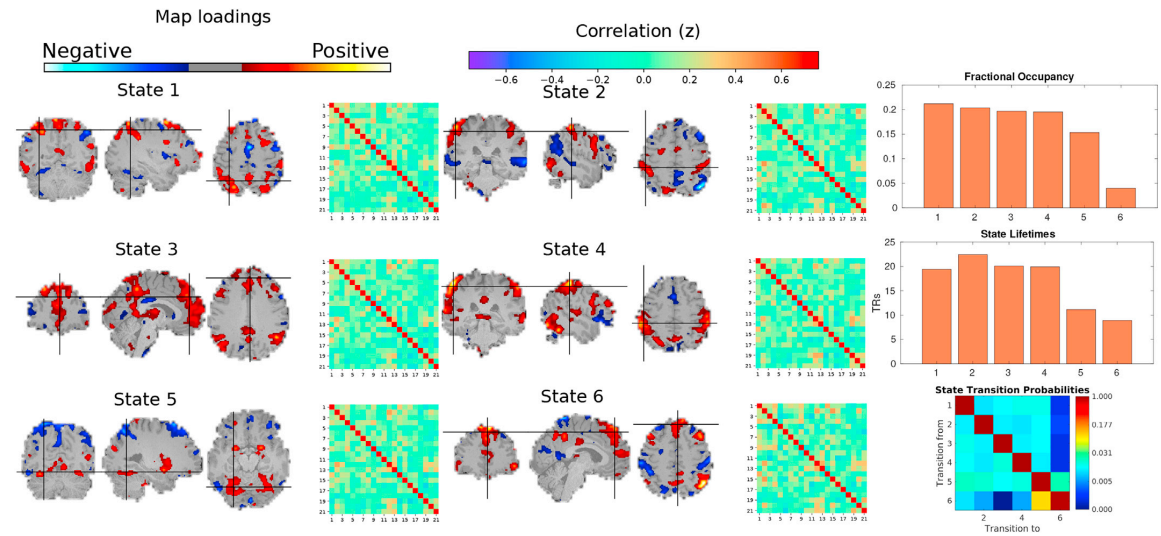


Fig. 9. Final VB-HMM VAR solution initialized with 7 states (one was emptied during training). The mean activity of each HMM-state is plotted, i.e. the mean of all time-points assigned to the same state. Furthermore, the empirical $p \times p$ correlation matrix for each state is also plotted (after Fisher transformation), where p is the number of ICs used (see appendix K). The states were sorted according to their fractional occupancy. Cut-coordinates were determined using Nilearn by finding the largest positive connected component after thresholding at the 95th percentile of the absolute values in the map. The fractional occupancy, mean lifetime and the transition probabilities between states is furthermore in the rightmost column.

VB-HMM VAR), by computing the mean activity of all the timepoints assigned to the same state. This is shown in Fig. 9 together with the FC-matrix pr. state, a bar plot of the fractional occupancy and mean lifetime (cf. appendix H) of each state. The states' spatial activity seems to resemble the default mode network (state 3 in particular) and the sensory motor network (state 2 and 4). We note that the states seem to have a mean lifetime in the range of 10–20 TR's (10–25 s) and that the transition matrix has a very diagonal structure indicating a lot of self-transitions, i.e. it is more likely to stay in the same state than jump to another state. Looking at the FC matrices all states seem to have a very diagonal structure with low variability over states.

For comparison, we ran the sliding window k-means approach (WKM) on the fMRI data with the same number of states as the VB-HMM-VAR estimated in the final run (i.e. 6 states). We used a tapered window with a window length of 22 TRs (corresponding to around 25s) sliding the window one TR at a time. We used the default MATLAB k++ initialization procedure (Arthur and Vassilvitskii, 2007) with Euclidean distance measure, and restarted the k-means procedure 100 times. However, we did not use ℓ_1-regularization as suggested in original WKM article (Allen et al., 2014), due to the well-posedness of the correlation matrices induced by the fairly low dimensionality of the problem.

For the WKM, the six states' mean activity, FC matrix and state characteristics are plotted in Fig. 10. The DMN activity seems to be separated over all the states. Looking at the mean lifetime of the states we see a very uniform distribution around 20 TRs, i.e. all states seems to have the same mean lifetime, which is probably mainly due to the window length. The FC notably varies more over states compared to the VB-HMM-VAR solution in Fig. 9.

Discussion

We have proposed a data-driven predictive framework for comparing and measuring generalization of dynamic functional connectivity (dFC) models. Using this framework we investigated a windowed covariance approach based on the infinite Wishart mixture model (IWMM) as well as the (window free) infinite HMMs (IHMM) specified by three different emission models (Nielsen et al., 2016; Baker et al., 2014; Vidaurre et al., 2016). We find that the extracted dynamics are heavily influenced by modeling assumptions. In synthetic data, where ground truth state sequences were available, it was clear that a misspecification of the model leads to an incorrect state sequence. Thus, we need to properly quantify how well certain model assumptions comply with the data observed. Here, the predictive assessment framework is able to quantify the number of states and appropriate emission model. We found the WKM to be robust towards model mismatch, however, we here in general have no a priori knowledge of either window length or the number of states that need to be specified. We found that the IWMM admits quantification of number of states within a WKM type of framework, but the choice of window length remains unresolved and influences results as illustrated in the synthetic study.

Hidden Markov models (HMMs) seem like a promising framework to circumvent the need to specify window lengths, learning state transitions and their smoothness as part of the inference. We considered both MCMC and variational Bayesian inference and consistently found that the choice of emission model heavily influences the identified functional dynamics and their interpretation as different emission models drive different dynamics. Our predictive framework admits quantification of the type of emission model that is most adequate for the system under consideration and our results points towards the vector autoregressive (VAR) model being a more flexible and better overall choice. It should be noted that in analysis of real data (EEG and fMRI) the data sets were lowpass and bandpass filtered respectively as part of the preprocessing, which may harm the estimated dynamics by driving the VAR-states towards characterizing properties of the preprocessing. In slowly fluctuating signals a large portion of the signal at time t can be explained by the signal at time $t-1$ which is exactly what the VAR(1)-model is doing in contrast to the other emission models (see also appendix section G). Preprocessing influences the estimated dynamics as shown in Hindriks et al. (2016). In this work we chose the default preprocessing pipelines as suggested by Wakeman and Henson (2015) and Poldrack et al. (2015), however we expect different preprocessing choices can favor different emission models. However, investigating these choices are out of scope of the current study. In our analyses of the EEG data (as well as in our synthetic study) there was a clear indication that the simpler HMMs (ZMG and

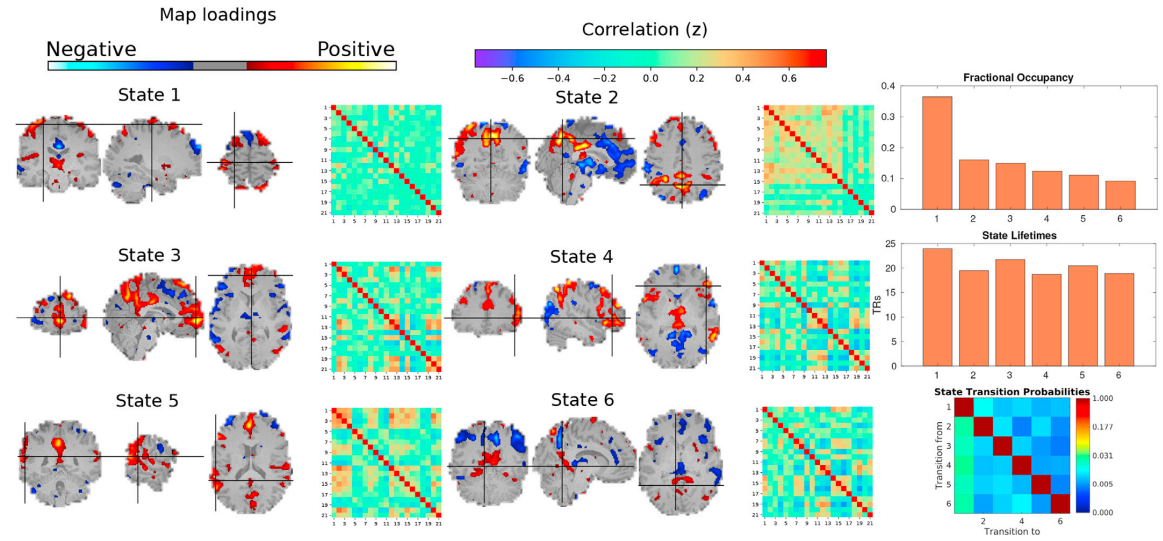


Fig. 10. Final WKM solution initialized with 6 states. The mean activity of each WKM-state, i.e. the mean of all time-points assigned to the same state, is plotted. Furthermore, the empirical $p \times p$ correlation matrix for each state is also plotted (after Fisher transformation), where p is the number of ICs used (see appendix K). The states were sorted according to their fractional occupancy prior to visualization. Cut-coordinates were determined using Nilearn by finding the largest positive connected component after thresholding at the 95th percentile of the absolute values in the map. The fractional occupancy, mean lifetime and the transition probabilities between states is furthermore in the rightmost column.

SSM) overestimated the number of states, whereas in the analyses of the fMRI data all emission models were more in agreement. We attribute this difference to the differences in signal-to-noise ratios and temporal resolutions, but note that this requires further investigation.

On the fMRI data we compared the HMM-VAR with the WKM visualizing the brain states extracted by the two frameworks (with the same number of states). It is clear that they find different brain state representations both in mean activity, FC and temporal characteristics. As such, the WKM finds more distinct states in terms of FC than the HMM-VAR. We attribute this to their different modeling assumptions, i.e. VB-HMM VAR is a model that generates data at the level of single time points whereas the WKM is driven by characterizing differences in the off-diagonal elements of the windowed covariance matrices. When looking at the lifetimes of the extracted states, all WKM states had approximately the same length dictated by the window length used, whereas the HMM-VAR due to its window-free approach estimated states with varying lifetime. This exemplifies that dynamics are driven by the underlying model assumptions. One could be tempted to interpret what the extracted states' represent in terms of brain function, however, the NMI results in Fig. 8a points toward issues with local minima in particular for the HMMs. We speculate that current dFC approaches are too flexible hampering the reliability (Choe et al., 2017), thus there seems to be a need for better inference procedures and constrained models promoting both reliability and generalization.

We compared two inference methods for the HMMs, namely Markov chain Monte Carlo (MCMC) in the form of the infinite hidden Markov model (IHMM) and variational Bayes hidden Markov model (VB-HMM). From a theoretical point of view the IHMM has the most desirable properties, i.e., we do not need to specify the number of states and we should obtain better estimates of the posterior distribution. However, in practice the IHMM and VB-HMM yield similar results, and if we factor in the computational complexity of the IHMM, the VB-HMM seems like the better choice in most practical applications.

Our results supports the conclusion that functional connectivity is best modeled using multiple states (Hutchison et al., 2013; Calhoun et al., 2014; Calhoun and Adali, 2016; Vidaurre et al., 2017b). In particular, our predictive assessment consistently finds support for functional neuroimaging data, i.e., fMRI and EEG data, are better accounted for by dynamic models (i.e., models having more than one state) which was consistently observed across models and data sets. As hypothesized we find that the more advanced HMM-VAR extracted fewer states than the simpler ZMG and SSM emission models. Thus, in theory a very complicated emission model (that we have not investigated here) could potentially capture everything as "one state".

There has recently been a lot of focus on null-models and stationarity in dFC (Zalesky and Breakspear, 2015; Laumann et al., 2016; Miller et al., 2017). For choosing an appropriate window-length in WKM the work of Zalesky and Breakspear (2015) provides some statistical analysis as to why the rule of thumb of 100 s windows from (Leonardi & Van De Ville, 2015) makes sense. Zalesky and Breakspear (2015) furthermore points out that the framework can detect changes in FC on shorter timescales (around 40 s); changes that can disappear if longer windows are used. Their conclusion also being that we need better generative null-models for dFC. While we do not claim that we have found the true null-model for dFC, we have demonstrated a framework that admits a comparison between models based on predictive likelihood. We compared the WKM with HMM-framework on fMRI data in qualitative way; however, since the WKM is not a model of data we cannot in an objective way compare the performance of the two models. Bzdok and Yeo (2017) argues that neuroscience is moving more and more towards out-of-sample generalization as an alternative to classical statistical inference and hypothesis testing, and we will argue that models of dFC will be more objectively comparable if they are generative and are able to extrapolate to held-out data. Importantly, the HMM is a generative model that contains the static model as a special case and by doing model order selection we test in a data-driven way whether or not the FC should be modeled static ($K = 1$) or dynamic ($K > 1$).

A very important point is that the proposed framework will only answer what model best explains the data at hand. To truly validate that the extracted dynamics correspond to neurophysiological mechanisms, we need more elaborate validation such as concurrent EEG-fMRI data or even invasive studies.

Our predictive assessment framework generalizes to arbitrary dynamic model specifications as long as a predictive likelihood can be calculated. For instance, the AR-order was fixed to one in this paper but could easily be learned using the framework presented (cf. appendix F). In this paper we also show two ways of using the predictive assessment framework promoting two different kinds of generalization, i.e. between-subject generalization and within-subject generalization. We are not claiming in any way that one should use one over the other, only that we have the power with this framework to investigate both types of generalization. The quantitative analysis of this paper points to dFC being heavily influenced by modeling assumptions and the proposed assessment provides a principled tool for future refinement and tailoring of models of dFC to better account for functional neuroimaging data.

Acknowledgements

Søren F.V. Nielsen, Mikkel N. Schmidt and Morten Mørup were supported by Lundbeckfonden (fellowship grant number R105-9813 to Morten Mørup). Kristoffer H. Madsen was supported by a Novo Nordisk Foundation Interdisciplinary Synergy Grant (grant number NNF14OC0011413).

Appendices.

A. Implementation details for IHMM

Both the IWMM and IHMMs were implemented using collapsed Gibbs sampling with split-merge proposals (Jain and Neal, 2004). α in the IWMM was inferred using random walk MCMC. The IHMMs were implemented on top of the MATLAB implementation made by Van Gael (2010), in which α and γ were sampled by placing vague Gamma priors on them. As pointed out in the literature (Van Gael et al., 2008) the Gibbs sampler has some mixing issues, so to overcome this we implemented a split-merge sampling procedure as described in (Jain and Neal, 2004) adapted to the IHMM framework. We use the same convention as in Van Gaels MATLAB-implementation namely that the first time point is assumed to have transitioned from state 1, i.e. $z_0 = 1$. Our MATLAB implementation is publicly available for download[8].

In all experiments, for both IHMM and VB-HMM, we fixed the AR-order in the VAR model to 1. In the IWMM and IHMM we parameterize the prior $\mathbf{\Sigma}_0 = \eta\mathbf{I}$. We found through experimentation that in most cases it is undesirable to infer the prior strength η, since it can yield a huge number of states.

[8] https://brainconnectivity.compute.dtu.dk/.

The prior strength acts a regularization on the number of states and should therefore be tuned in order for the model to best characterize test data. We therefore learned this parameter using cross-validation considering values in the range $\eta \in [10^{\log\sigma-5}, 10^{\log\sigma+5}]$, where σ is the scale of the data (sampled equidistantly in the log-domain). Note that the most computationally demanding operation in the inference is the calculation of the determinant of a matrix representing the sufficient statistic for each state. This can in the case of the ZMG and SSM emission-models be handled efficiently using Cholesky-factorizations, which makes the algorithm scale as follows; for a particular iteration with K states on a p dimensional dataset of length T the computational cost is $O(TKp^2)$. For the VAR-emission the Cholesky-trick cannot be applied and thus the computational cost scales as $O(TK(pr)^3)$, where r is the lag of the VAR-model.

B. Variational Bayes hidden Markov model

In this paper we use the (finite) variational Bayesian HMM (VB-HMM) implementation from (50), where the generative model (without specifying the emission distribution) can be written as,

$$\boldsymbol{\pi}_0 \sim \mathrm{Dir}(\boldsymbol{\kappa}) \tag{S.1}$$

$$\boldsymbol{\pi}^{(k)} \sim \mathrm{Dir}\left(\boldsymbol{\lambda}^{(k)}\right), \tag{S.2}$$

$$z_t \left| z_{t-1} \right. \sim \mathrm{Multinomial}\left(\boldsymbol{\pi}^{(z_{t-1})}\right), \tag{S.3}$$

$$\theta^{(k)} \sim H \tag{S.4}$$

$$\mathbf{x}_t \sim F\left(\theta^{(z_t)}\right), \tag{S.5}$$

in which $\boldsymbol{\pi}_0$ is the initial state distribution vector (size K), Dir() is the Dirichlet distribution, $\boldsymbol{\kappa}$ is the prior vector for the initial distribution, $\boldsymbol{\pi}^{(k)}$ is a row of the transition matrix, $\boldsymbol{\lambda}^{(k)}$ is the associated prior to that row, z_t is the integer valued state taking possible values from $1..K$ at time point t, $\theta^{(k)}$ are all state relevant parameters drawn from the unknown prior $H(\cdot)$ for state k, and $\mathbf{x}_t$ is the observation at time t with emission distribution $F(\cdot)$. The graphical model for a probabilistic HMM with unspecified emission distribution (more on this in section 2.2.1) can be seen in Figure S1a. Inference in the model is done using the standard variational Bayes (VB) update rules (Rezek & Roberts, 2005), where each part of the graphical model is updated in turn. For a K-state model run on a p-dimensional dataset with T time-points computationally the algorithm scales as follows; the ZMG and SSM emission-models scale as $O(TKp^2)$ and the VAR emission-model as $O(TK(pr)^3))$ both due to a matrix inversion. However, a lot these calculations are highly parallelizeable making the VB-HMM much faster in practice compared to the IHMM. The graphical model can be seen in Figure S1a.

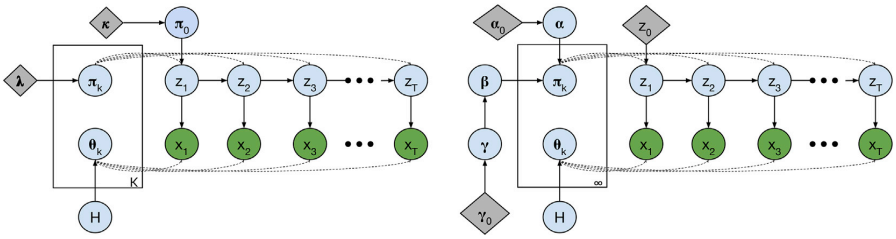


Fig. S.1. Graphical model for the two Bayesian hidden Markov models used in this paper. All blue circles are estimated in the inference procedure, green circles are observed and grey squares are parameters we fix. We observe the p-dimensional time series $\mathbf{x}_t$, which are dependent on each other through the 1st order Markovian hidden variable z_t. The transition probability between states is modeled in the transition matrix $\boldsymbol{\pi}$. Each state has some associated state-specific parameters $\theta^{(k)}$, with an unspecified prior distribution, H.

B.1. Predictive likelihood in VB-HMM

Let $\boldsymbol{\theta}_{obs}$ denote all emission-parameters. For the VB-HMM we make use of the variational posterior $Q_X(\boldsymbol{\theta}_{obs})$, $Q_X(\boldsymbol{\pi}_0)$, and $Q_X(\boldsymbol{\pi})$ which has been fitted to the training data, and furthermore bound this approximation (using Jensens inequality) by performing an expectation step on the state sequence of the test data, fixing all other parameters in the model, except the $Q_{X^*}(\mathbf{z}^*)$ distribution. This yields the log predictive likelihood,

$$\begin{aligned} &\ln p(\mathbf{X}^*|\mathbf{X}) \approx \ln \int\int\int\int [p(\mathbf{X}^*\mathbf{z}^*|\boldsymbol{\pi}_0, \boldsymbol{\pi}, \boldsymbol{\theta}_{obs}) Q_X(\boldsymbol{\pi}_0) Q_X(\boldsymbol{\pi}) Q_X(\boldsymbol{\theta}_{obs})] d\boldsymbol{\pi}_0 d\boldsymbol{\pi} d\boldsymbol{\theta}_{obs} d\mathbf{z}^* \\ &\geq \langle \ln p(\mathbf{X}^*, \mathbf{z}^*|\boldsymbol{\pi}_0, \boldsymbol{\pi}, \boldsymbol{\theta}_{obs}) \rangle_{Q_X(\boldsymbol{\pi}_0) Q_X(\boldsymbol{\pi}) Q_X(\boldsymbol{\theta}_{obs}) Q_{X^*}(\mathbf{z}^*)} \\ &- \langle \ln Q_{X^*}(\mathbf{z}^*) \rangle_{Q_{X^*}(\mathbf{z}^*)}, \end{aligned} \tag{S.6}$$

in which $\mathbf{z}^*$ is the state sequence of the test set. This is equivalent to estimating the free-energy (Vidaurre et al. (2016)) on the test set, i.e., without updating $Q(\boldsymbol{\pi})$, $Q(\boldsymbol{\pi}_0)$, and $Q(\boldsymbol{\theta}_{obs})$ and not including terms in the free-energy that have not changed compared to the free-energy of the training set.

C. Predictive likelihood in IWMM

In the case of the IWMM, we have conjugacy between the training-posterior $p(\boldsymbol{\Theta}|\mathscr{C})$, in which $\mathscr{C}$ is the collection of all training data scatter matrices, and the likelihood function $p(\mathbf{C}^*|\boldsymbol{\Theta})$ if we condition on the state sequence of the training data, $\mathbf{z}$. Using samples of $\mathbf{z}$ during the MCMC sampling

procedure, $\mathbf{z}^{(t)}$, we can approximate the predictive likelihood as,

$$p(\boldsymbol{C}^*|\mathscr{C}) \approx \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K+1}\frac{N_k^{(t)}}{N+\alpha^{(t)}}\int p(\boldsymbol{C}^*|\boldsymbol{\Sigma}^{(k)})p(\boldsymbol{\Sigma}^{(k)}|\mathscr{C},\ \mathbf{z}^{(t)},\ \eta^{(t)})d\boldsymbol{\Sigma}^{(k)}, \tag{S.7}$$

where $N_{K+1}^{(t)} = \alpha^{(t)}$, $N_k^{(t)}$ is the number of time-points in $\mathbf{z}^{(t)}$ assigned to cluster k, and N is the total number of time-points. Due the aforementioned conjugacy we integrate out $\boldsymbol{\Sigma}^{(k)}$ analytically from the predictive likelihood in the integral above.

D. Predictive likelihood in the IHMM

In the IHMM we obtain samples of the transition matrix $\boldsymbol{\pi}$ and $\boldsymbol{\theta}_{obs}$ during the MCMC sampling procedure, enabling us to integrate out those parameters using standard MCMC integration. This yields the log predictive likelihood estimate using T samples,

$$\ln p(\mathbf{X}^*|\mathbf{X}) \approx \ln\frac{1}{T}\sum_{t=1}^{T}\sum_{\mathbf{z}'}p\left(\mathbf{X}^*, \mathbf{z}'\middle|\boldsymbol{\pi}^{(t)}, \boldsymbol{\theta}_{obs}^{(t)}\right). \tag{S.8}$$

Here we analytically sum over all possible state sequences $\mathbf{z}'$, assuming that the found number of states is correct. This can be done efficiently using dynamic programming (Viterbi, 1967).

E. Synthetic study of VB-HMM

We demonstrate on synthetic data with three states from each of the emission models how the VB-HMM models perform. For each model we test on a hold-out validation set what number of states in the model yields the best predictive likelihood. In Figure S2, we show the estimated state sequences for each emission model and data set for the "cross"-validated number of states on the concatenated training and validation set. As with the IHMM we note that the simpler models (ZMG and SSM) struggle on data from the more complex emission model (VAR), whereas the VAR-model can adapt to the simple data.

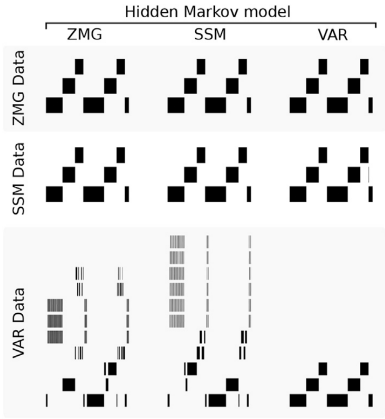


Fig. S.2. Estimated state sequences for synthetic data generated from hidden Markov models. Top: Zero mean Gaussian (ZMG) emission. Middle: State-specific mean (SSM) emissions. Bottom: Vector autoregressive (VAR) emission. Results are shown for data generated according to the hidden Markov models with ZMG, SSM, and VAR emission models fitted using variational Bayes. The true state sequence is shown in Fig. 3.

F. Selection of the VAR-order using predictive likelihood

The order of the autoregressive mean, r, that we use in the IHMM-VAR and VB-HMM-VAR is an important parameter, and how to choose this is still unclear. Our predictive likelihood framework also offers the option to estimate the optimal r to use. We tested this in a synthetic experiment where we used the VAR-data from section 3.2, with three states with state-specific VAR-coefficients, each of order one ($r = 1$). Then we fitted the IHMM-VAR and the VB-HMM-VAR using different VAR-orders from $r = 1..5$ on the training data. We furthermore ran the VB-inference for different number of states $K = 1, 2, 3$. The predictive results on the test data can be seen in Figure S3. For the IHMM-VAR model we see that the predictive log Bayes factor decreases as we increase r, correctly identifying the order to be $r = 1$. In the case of the VB-HMM-VAR, we see that if we use the wrong number of states (i.e. $K = 1, 2$), the predictive framework favors using higher model orders, whereas when we use the correct number of states $K = 3$ the framework correctly points toward model order $r = 1$. This brings up the discussion of how model order and number of states together affect our interpretation of dynamics. However, in most cases we find it appropriate to use an order of one (cf. discussion section 4 for more details on this).

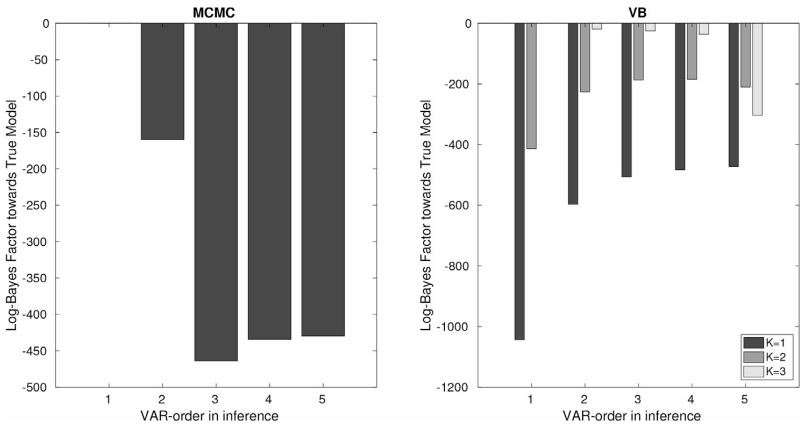


Fig. S.3. Log Bayes factor on test set vs the VAR-order one model for different orders. In case of the VB-HMM-VAR all Bayes factors are towards the correct model (i.e. $K = 3$ and $r = 1$). We generated a training and test data from a three state model, where each state had a VAR-emission of order 1. We then trained the IHMM-VAR and VB-HMM-VAR using different orders and number of states (for VB), and calculated predictive likelihood on the test set.

G. HMM: Synthetic study with fMRI signal properties

We investigated the influence of noise in the data together with more realistic fMRI signal properties. The synthetic data were generated by first sampling $p = 5$ random independent components (IC) from the resting-state fMRI data (see section 3.4) out of the 21 ICs that were deemed neural. Then we estimated the covariance matrix from the first 25 time points using only p ICs of three randomly sampled sessions from the training data, and used these as three ground truth functional connectivity (FC) states. We estimated the power-spectrum from a single session and generated three data sets (training, validation and test) by first generating random data preserving the estimated power-spectrum and then introducing systematic coupling using the three estimated FC states. Finally, we added a level of white noise to obtain data with a specific SNR. We did this for SNR $= [-6, 6]$ dB and repeated the data generation process 10 times. Figure S4 shows the mean predictive log likelihood of each of the VB-HMMs on the test set, and the normalized mutual information towards the true state sequence; in both cases after optimizing the number of states using the validation set. We see that the three models perform very similarly in terms of predictive performance on the held-out data, with the VAR slightly ahead in the high SNR regime. We attribute this to the smoothness of the data induced by preprocessing of the fMRI data. In terms of finding the true state sequence the VAR and ZMG follow each other closely but the ZMG breaks off and outperforms the two other models at around SNR $= 0$. This can be explained by VAR being able to characterize the power-spectrum better in the high SNR regime; and as the SNR decreases, the power-spectrum is destroyed by the white noise making it easier for the ZMG to find the underlying state sequence.

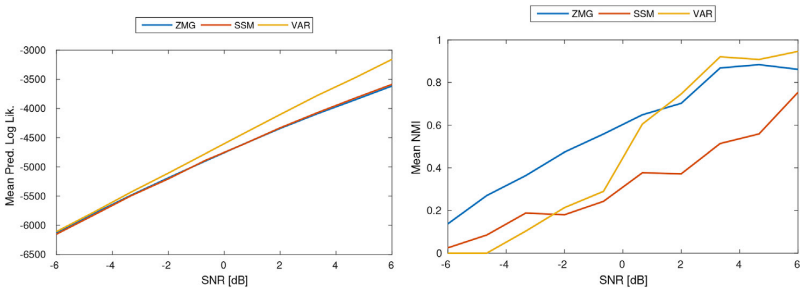


Fig. S.4. Results from synthetic analysis with fMRI signal properties. The above results are averages over 10 data sets.

H. HMM: Summary statistics

We use two summary statistics in the paper to quantify the characteristics of the extracted states, namely *fractional occupancy* and *mean lifetime* as defined in (Baker et al., 2014).

Fractional occupancy

The fractional occupancy, f_k, of each state is the empirical estimate of the probability of being in this state at any point in time. It is defined for a given state sequence $\boldsymbol{z}$ of length T as,

$$f_k = \frac{\sum_t \delta(z_t = k)}{T}, \tag{S.9}$$

where $\delta(z_t = k)$ is the delta function that takes on the value 1 if z_t is equal to k and is zero otherwise.

Mean lifetime

The mean lifetime, ml_k, is an empirical estimate of how long we expect a certain state to persist. It is defined as,

$$ml_k = \frac{\sum_t \delta(z_t = k)}{\sum_t \delta(z_t = k) \cdot \delta(z_{t-1} \neq k)}, \tag{S.10}$$

in which $\delta(z_t \neq k)$ is the delta function that takes on value 1 if z_t is not equal to k and zero otherwise.

I. HMM: Robustness of the inference procedures

To investigate how the different states are populated over restarts and emission model in the HMM-framework we show the empirical state-sequence distribution for the two real-world data sets.

I.1. EEG: face scrambling famous condition

The fractional occupancy of each state (ordered by magnitude) is shown as a stacked bar plot in Figure S5. The ZMG and SSM employed more states to explain the data compared to the VAR emission model. Comparing results between the IHMM (using MCMC inference) and the VB-HMM (using variational inference), the two inference methods identified the same pattern, namely that the VAR found fewer states than the two simpler emission models. Both inference procedures found fairly consistent state occupancy distributions over multiple restarts. Looking at the different parameterisations, the estimated dFC dynamics were heavily influenced by the choice of emission model.

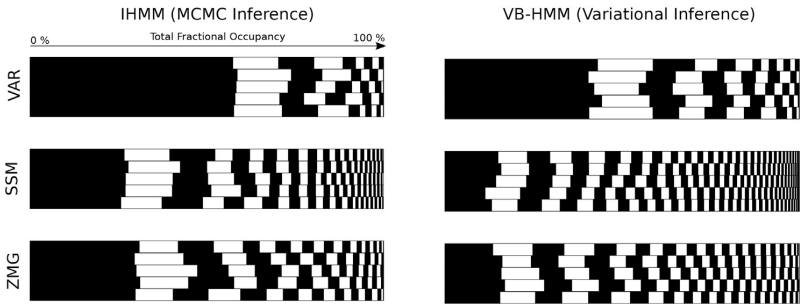


Fig. S.5. Fractional occupancy of each state for each model over 5 restarts, when trained on the first condition *famous* from the EEG data. The states are shown as a stacked bar plot ordered by their fractional occupancy and alternately colored black and white.

I.2. fMRI: single subject resting-state

The fractional occupancy of each state for each emission model and restart can be seen in the stacked bar plot in Figure S6. The VAR model consistently found six states, whereas the ZMG and SSM found 7 and 8 states respectively. The occupancy of each states was fairly robust over restarts in all emission models.

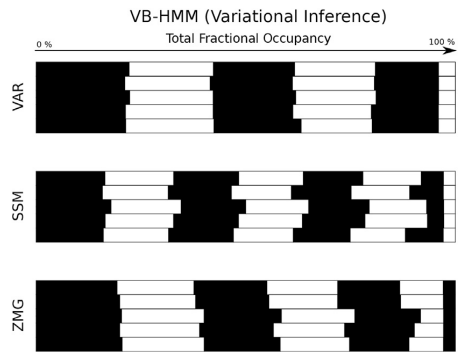


Fig. S.6. Fractional occupancy (fMRI resting state data) of each state for each model over 5 restarts (on the training data). The states are shown as a stacked bar plot ordered by their fractional occupancy and alternately colored black and white.

J. HMM on EEG-data: sampling the posterior distribution

We illustrate what the three different IHMM-emission models have learned on the first condition (*famous*) from the EEG-data Wakeman and Henson (2015). Figure S7 shows the estimated state sequence for the first subject in the first condition for each of the models, and illustrates data sampled from the fitted posterior distributions. All models divided the ERP into a number of states: The ZMG and SSM models found more states than the VAR model, and the data sampled from the posterior of the ZMG and SSM models did not reflect the smoothness of the true ERP response (see Fig. 7c). The VAR model found a state sequence that was in better correspondence with the ERP response compared to the other models, including a baseline state that appears before and after the ERP.

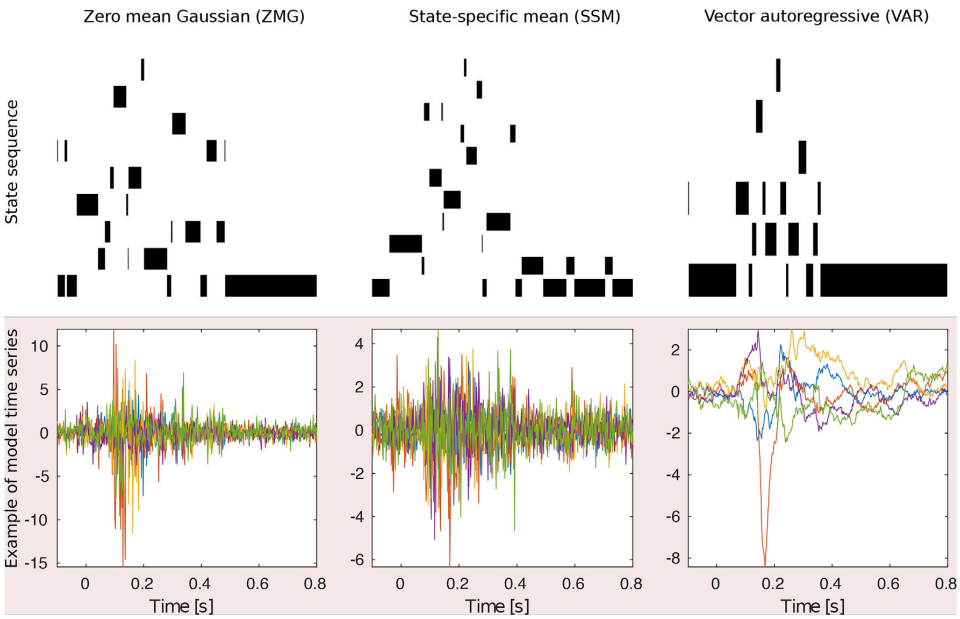


Fig. S.7. Estimated state sequences (EEG data) for the first subject and first condition are shown for all the emission models of the IHMM. Furthermore, we show data sampled using the posterior parameters obtained from the last sample of the first MCMC chain.

K. Resting state fMRI data: group ICA components

In this section we plot the spatial maps of the group independent components estimated as described in the results section 3.4. They can be seen in Figure S8 in three views chosen using the plot_stat_map function from Nilearn[9]. The threshold was chosen to be the 95th-percentile of the absolute values in the image.

[9] http://nilearn.github.io/.

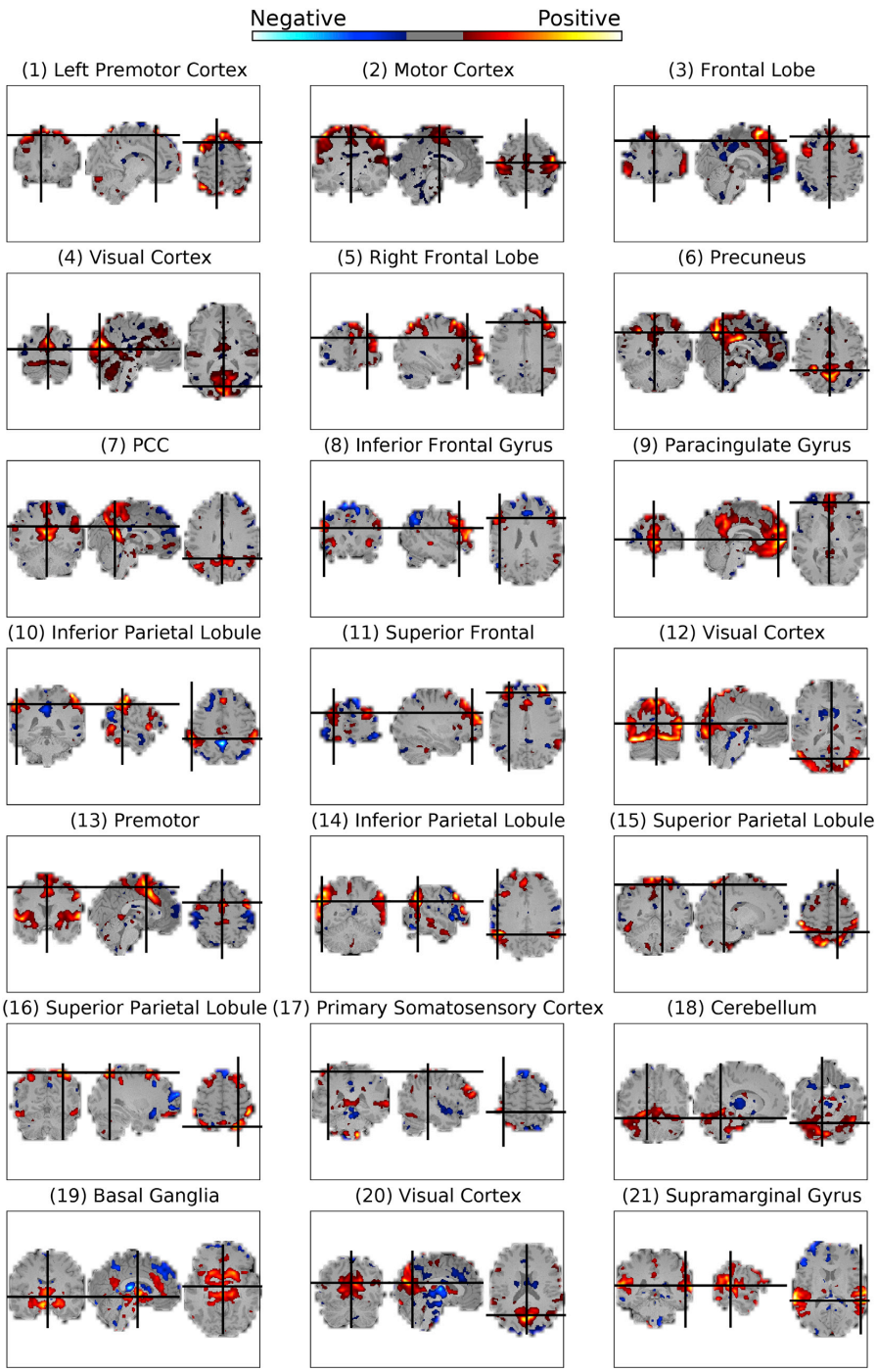


Fig. S.8. Spatial maps of the retained 21 ICA components from the resting state fMRI data analyzed in this paper. Cut-coordinates were determined using Nilearn by finding the largest positive connected component after thresholding at the 95th percentile of the absolute values in the map.

References

Abrahamsen, T.J., Hansen, L.K., 2011. A cure for variance inflation in high dimensional kernel principal component analysis. J. Mach. Learn. Res. 12, 2027–2044.

Aldous, D.J., 1985. Exchangeability and related topics. In: Hennequin, P.L. (Ed.), École d'Été de Probabilités de Saint-Flour XIII — 1983 Lecture Notes in Mathematics. Springer Berlin Heidelberg, pp. 1–198. https://doi.org/10.1007/BFb0099421.

Allen, E.A., Damaraju, E., Plis, S.M., Erhardt, E.B., Eichele, T., Calhoun, V.D., 2014. Tracking whole-brain connectivity dynamics in the resting state. Cerebr. Cortex 24, 663–676. https://doi.org/10.1093/cercor/bhs352.

Arthur, D., Vassilvitskii, S., 2007. k-means++: the advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (Pp. 1027–1035).

Baker, A.P., Brookes, M.J., Rezek, I.A., Smith, S.M., Behrens, T., others, 2014. Fast transient networks in spontaneous human brain activity. eLife 3, e01867. https://doi.org/10.7554/eLife.01867.001.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J.W., Hasson, U., Norman, K.A., 2017. Discovering event structure in continuous narrative perception and memory. Neuron 95, 709–721. https://doi.org/10.1016/j.neuron.2017.06.041 e5.

Beal, M.J., 2003. Variational Algorithms for Approximate Bayesian Inference. Ph.D. thesis.

Beal, M.J., Ghahramani, Z., Rasmussen, C.E., 2002. The infinite hidden markov model. In: Dietterich, T.G., Becker, S., Ghahramani, Z. (Eds.), Advances in Neural Information Processing Systems 14. MIT Press, pp. 577–584.

Bell, A.J., Sejnowski, T.J., 1995. An information-maximization approach to blind separation and blind deconvolution. Neural Comput. 7, 1129–1159.

Blei, D.M., Jordan, M.I., 2006. Variational inference for dirichlet process mixtures. Bayesian Anal. 1, 121–143. https://doi.org/10.1214/06-BA104.

Bzdok, D., Yeo, B.T.T., 2017. Inference in the age of big data: future perspectives on neuroscience. Neuroimage. https://doi.org/10.1016/j.neuroimage.2017.04.061.

Calhoun, V.D., Adali, T., 2016. Time-Varying brain connectivity in fMRI data: whole-brain data-driven approaches for capturing and characterizing dynamic states. IEEE Signal Process. Mag. 33, 52–66. https://doi.org/10.1109/MSP.2015.2478915.

Calhoun, V.D., Adali, T., Pearlson, G.D., Pekar, J.J., 2001. A method for making group inferences from functional MRI data using independent component analysis. Hum. Brain Mapp. 14, 140–151.

Calhoun, V.D., Miller, R., Pearlson, G., Adalı, T., 2014. The chronnectome: time-varying connectivity networks as the next frontier in fMRI data discovery. Neuron 84, 262–274. https://doi.org/10.1016/j.neuron.2014.10.015.

Cherian, A., Morellas, V., Papanikolopoulos, N., 2016. Bayesian nonparametric clustering for positive definite matrices. IEEE Trans. Pattern Anal. Mach. Intell. 38, 862–874. https://doi.org/10.1109/TPAMI.2015.2456903.

Choe, A.S., Nebel, M.B., Barber, A.D., Cohen, J.R., Xu, Y., Pekar, J.J., Caffo, B., Lindquist, M.A., 2017. Comparing test-retest reliability of dynamic functional connectivity methods. Neuroimage 158, 155–175. https://doi.org/10.1016/j.neuroimage.2017.07.005.

Du, W., Calhoun, V.D., Li, H., Ma, S., Eichele, T., Kiehl, K.A., Pearlson, G.D., Adali, T., 2012. High classification accuracy for schizophrenia with rest and task FMRI data. Front. Hum. Neurosci. 6, 145. https://doi.org/10.3389/fnhum.2012.00145.

Ferguson, T.S., 1973. A bayesian analysis of some nonparametric problems. Ann. Stat. 1, 209–230.

Finn, E.S., Shen, X., Scheinost, D., Rosenberg, M.D., Huang, J., Chun, M.M., Papademetris, X., Constable, R.T., 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nat. Neurosci. 18, 1664–1671. https://doi.org/10.1038/nn.4135.

Fox, E., Sudderth, E.B., Jordan, M.I., Willsky, A., 2011. Bayesian nonparametric inference of switching dynamic linear models. Signal Processing, IEEE Transactions on 59, 1569–1585. https://doi.org/10.1109/TSP.2010.2102756.

Hidot, S., Saint-Jean, C., 2010. An Expectation–Maximization algorithm for the wishart mixture model: application to movement clustering. Pattern Recogn. Lett. 31, 2318–2324. https://doi.org/10.1016/j.patrec.2010.07.002.

Hindriks, R., Adhikari, M.H., Murayama, Y., Ganzetti, M., Mantini, D., Logothetis, N.K., Deco, G., 2016. Can sliding-window correlations reveal dynamic functional connectivity in resting-state fMRI? Neuroimage 127, 242–256. https://doi.org/10.1016/j.neuroimage.2015.11.055.

Horn, J.L., 1965. A rationale and test for the number of factors in factor analysis. Psychometrika 30, 179–185.

Hutchison, R.M., Womelsdorf, T., Allen, E.A., Bandettini, P.A., Calhoun, V.D., Corbetta, M., Della Penna, S., Duyn, J.H., Glover, G.H., Gonzalez-Castillo, J., Handwerker, D.A., Keilholz, S., Kiviniemi, V., Leopold, D.A., de Pasquale, F., Sporns, O., Walter, M., Chang, C., 2013. Dynamic functional connectivity: promise, issues, and interpretations. Neuroimage 80, 360–378. https://doi.org/10.1016/j.neuroimage.2013.05.079.

Jain, S., Neal, R.M., 2004. A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. J. Comput. Graph Stat. 13, 158–182. https://doi.org/10.1198/1061860043001.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90, 773–795. https://doi.org/10.1080/01621459.1995.10476572.

Khanna, A., Pascual-Leone, A., Michel, C.M., Farzan, F., 2015. Microstates in resting-state EEG: current status and future directions. Neurosci. Biobehav. Rev. 49, 105–113. https://doi.org/10.1016/j.neubiorev.2014.12.010.

Korzen, J., Madsen, K.H., Mørup, M., 2014. Quantifying temporal states in rs-fMRI data using bayesian nonparametrics. Poster presentation at Human Brain Mapping 2014.

Laumann, T.O., Snyder, A.Z., Mitra, A., Gordon, E.M., Gratton, C., Adeyemo, B., Gilmore, A.W., Nelson, S.M., Berg, J.J., Greene, D.J., McCarthy, J.E., Tagliazucchi, E., Laufs, H., Schlaggar, B.L., Dosenbach, N.U.F., Petersen, S.E., 2016. On the stability of BOLD fMRI correlations. Cerebr. Cortex. https://doi.org/10.1093/cercor/bhw265.

Leonardi, N., Van De Ville, D., 2015. On spurious and real fluctuations of dynamic functional connectivity during rest. Neuroimage 104, 430–436. https://doi.org/10.1016/j.neuroimage.2014.09.007.

Miller, R.L., Adali, T., Levin-Schwartz, Y., Calhoun, V.D., 2017. Resting-state FMRI Dynamics and Null Models: Perspectives, Sampling Variability, and Simulations. https://doi.org/10.1101/153411.

Murray, M.M., Brunet, D., Michel, C.M., 2008. Topographic ERP analyses: a step-by-step tutorial review. Brain Topogr. 20, 249. https://doi.org/10.1007/s10548-008-0054-5.

Nielsen, S.F.V., Madsen, K.H., Røge, R., Schmidt, M.N., Mørup, M., 2016. Nonparametric modeling of dynamic functional connectivity in fMRI data. In: Rish, I., Wehbe, L., Langs, G., Grosse-Wentrup, M., Murphy, B., Cecchi, G. (Eds.), NIPS 2015 Workshop on Machine Learning and Interpretation in Neuroimaging. arxiv.org.

Nielsen, S.F.V., Madsen, K.H., Schmidt, M.N., Mørup, M., 2017. Modeling dynamic functional connectivity using a wishart mixture model. In: 2017 International Workshop on Pattern Recognition in Neuroimaging (PRNI). IEEE, pp. 1–4. https://doi.org/10.1109/PRNI.2017.7981505.

O'Neill, G.C., Tewarie, P., Vidaurre, D., Liuzzi, L., Woolrich, M.W., Brookes, M.J., 2017. Dynamics of large-scale electrophysiological networks: a technical review. Neuroimage. https://doi.org/10.1016/j.neuroimage.2017.10.003.

Orbanz, P., Teh, Y.W., 2011. Bayesian nonparametric models. In: Sammut, C., Webb, G.I. (Eds.), Encyclopedia of Machine Learning. Springer US, pp. 81–89. https://doi.org/10.1007/978-0-387-30164-8_66.

Ott, C.G.M., Langer, N., Oechslin, M.S., Meyer, M., Jäncke, L., 2011. Processing of voiced and unvoiced acoustic stimuli in musicians. Front. Psychol. 2, 195. https://doi.org/10.3389/fpsyg.2011.00195.

Patel, A.X., Kundu, P., Rubinov, M., Jones, P.S., Vértes, P.E., Ersche, K.D., Suckling, J., Bullmore, E.T., 2014. A wavelet method for modeling and despiking motion artifacts from resting-state fMRI time series. Neuroimage 95, 287–304. https://doi.org/10.1016/j.neuroimage.2014.03.012.

Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004. Comparing dynamic causal models. Neuroimage 22, 1157–1172. https://doi.org/10.1016/j.neuroimage.2004.03.026.

Poldrack, R.A., Laumann, T.O., Koyejo, O., Gregory, B., Hover, A., Chen, M.-Y., Gorgolewski, K.J., Luci, J., Joo, S.J., Boyd, R.L., Hunicke-Smith, S., Simpson, Z.B., Caven, T., Sochat, V., Shine, J.M., Gordon, E., Snyder, A.Z., Adeyemo, B., Petersen, S.E., Glahn, D.C., Reese Mckay, D., Curran, J.E., Göring, H.H.H., Carless, M.A., Blangero, J., Dougherty, R., Leemans, A., Handwerker, D.A., Frick, L., Marcotte, E.M., Mumford, J.A., 2015. Long-term neural and physiological phenotyping of a single human. Nat. Commun. 6, 8885. https://doi.org/10.1038/ncomms9885.

Rashid, B., Arbabshirani, M.R., Damaraju, E., Cetin, M.S., Miller, R., Pearlson, G.D., Calhoun, V.D., 2016. Classification of schizophrenia and bipolar patients using static and dynamic resting-state fMRI brain connectivity. Neuroimage 134, 645–657. https://doi.org/10.1016/j.neuroimage.2016.04.051.

Rasmussen, C.E., 1999. The infinite gaussian mixture model. In: Advances in Neural Information Processing Systems 12. kyb.tue.mpg.de.

Rezek, I., Roberts, S., 2005. Ensemble hidden markov models with extended observation densities for biosignal analysis. In: Dirk Husmeier, M., Richard Dybowski, M., Roberts, Stephen (Eds.), Probabilistic Modeling in Bioinformatics and Medical Informatics Advanced Information and Knowledge Processing. Springer London, pp. 419–450. https://doi.org/10.1007/1-84628-119-9_14.

Ryali, S., Supekar, K., Chen, T., Kochalka, J., Cai, W., Nicholas, J., Padmanabhan, A., Menon, V., 2016. Temporal dynamics and developmental maturation of salience, default and Central-Executive network interactions revealed by variational bayes hidden markov modeling. PLoS Comput. Biol. 12, e1005138 https://doi.org/10.1371/journal.pcbi.1005138.

Shakil, S., Lee, C.-H., Keilholz, S.D., 2016. Evaluation of sliding window correlation performance for characterizing dynamic functional connectivity and brain states. Neuroimage 133, 111–128. https://doi.org/10.1016/j.neuroimage.2016.02.074.

Van Gael, J., 2010. The Infinite Hidden Markov Model 0.5. http://mloss.org/software/view/205/.

Van Gael, J., 2011. Bayesian Nonparametric Hidden Markov Models. Ph.D. thesis. University of Cambridge.

Van Gael, J., Saatci, Y., Teh, Y.W., Ghahramani, Z., 2008. Beam sampling for the infinite hidden markov model. In: Proceedings of the 25th International Conference on Machine Learning ICML '08. ACM, New York, NY, USA, pp. 1088–1095. https://doi.org/10.1145/1390156.1390293.

Vidaurre, D., Abeysuriya, R., Becker, R., Quinn, A.J., Alfaro-Almagro, F., Smith, S.M., Woolrich, M.W., 2017a. Discovering dynamic brain networks from big data in rest and task. Neuroimage. https://doi.org/10.1016/j.neuroimage.2017.06.077.

Vidaurre, D., Quinn, A.J., Baker, A.P., Dupret, D., Tejero-Cantero, A., Woolrich, M.W., 2016. Spectrally resolved fast transient brain states in electrophysiological data. Neuroimage 126, 81–95. https://doi.org/10.1016/j.neuroimage.2015.11.047.

Vidaurre, D., Smith, S.M., Woolrich, M.W., 2017b. Brain network dynamics are hierarchically organized in time. Proc. Natl. Acad. Sci. U. S. A.,. https://doi.org/10.1073/pnas.1705120114.

Viterbi, A., 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans. Inf. Theor. 13, 260–269. https://doi.org/10.1109/TIT.1967.1054010.

Wakeman, D.G., Henson, R.N., 2015. A multi-subject, multi-modal human neuroimaging dataset. Sci. Digest 2, 150001. https://doi.org/10.1038/sdata.2015.1.
Zalesky, A., Breakspear, M., 2015. Towards a statistical test for functional connectivity dynamics. Neuroimage 114, 466–470. https://doi.org/10.1016/j.neuroimage.2015.03.047.
Zalesky, A., Fornito, A., Cocchi, L., Gollo, L.L., Breakspear, M., 2014. Time-resolved resting-state brain networks. Proc. Natl. Acad. Sci. U. S. A. 111, 10341–10346. https://doi.org/10.1073/pnas.1400181111.

A.4 Evaluating Models of Dynamic Functional Connectivity using Predictive Classification Accuracy

Nielsen, Søren F V, Yuri Levin-Schwartz, Diego Vidaurre, Tulay Adali, Vince D Calhoun, Kristoffer H Madsen, Lars Kai Hansen, and Morten Mørup (2018). "Evaluating models of dynamic functional connectivity using predictive classification accuracy". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Calgary, AB, Canada.

EVALUATING MODELS OF DYNAMIC FUNCTIONAL CONNECTIVITY USING PREDICTIVE CLASSIFICATION ACCURACY

Søren Føns Vind Nielsen[1], Yuri Levin-Schwartz[2], Diego Vidaurre[3], Tulay Adali[2], Vince D. Calhoun[5,6], Kristoffer H. Madsen[1,4], Lars Kai Hansen[1] and Morten Mørup[1]

[1] Department of Applied Mathematics and Computer Science, Technical University of Denmark
[2] Department of CSEE, University of Maryland, Baltimore County, USA
[3] OHBA, Welcome Centre for Integrative Neuroimaging, Department of Psychiatry, University of Oxford, UK
[4] Danish Research Centre for Magnetic Resonance, Copenhagen University Hospital Hvidovre, Denmark
[5] The Mind Research Network, Albuquerque, USA
[6] Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, USA

ABSTRACT

Dynamic functional connectivity has become a prominent approach for tracking the changes of macroscale statistical dependencies between regions in the brain. Effective parametrization of these statistical dependencies, referred to as brain states, is however still an open problem. We investigate different emission models in the hidden Markov model framework, each representing certain assumptions about dynamic changes in the brain. We evaluate each model by how well they can discriminate between schizophrenic patients and healthy controls based on a group independent component analysis of resting-state functional magnetic resonance imaging data. We find that simple emission models without full covariance matrices can achieve similar classification results as the models with more parameters. This raises questions about the predictability of dynamic functional connectivity in comparison to simpler dynamic features when used as biomarkers. However, we must stress that there is a distinction between characterization and classification, which has to be investigated further.

Index Terms— Dynamic functional connectivity, Hidden Markov models, Classification, Schizophrenia

1. INTRODUCTION

In the study of how the brain integrates information, communication between disjoint regions is often described using *functional connectivity* (FC). Over the last two decades, FC analysis has relied on a stationary assumption, i.e. that the statistical dependencies between regions do not change over time. This assumption has been shown to disregard a potential wealth of information in the changes in between-region connectivity, especially in resting-state functional magnetic resonance imaging (rs-fMRI) where this analysis approach has been coined dynamic functional connectivity (dFC) [1, 2].

The most widely used approach in the dFC litterature is the sliding-window correlation (SWC) [3], in which the (regularized) correlation matrix was estimated in windows slid one time-step at a time on group independent component analysis (gICA) time-courses from rs-fMRI from healthy subjects. After applying a k-means clustering to the estimated windowed correlation matrices they found that the seven clusters extracted varied especially in their connectivity within the default mode network.

However, SWC has been criticized because the choice of window-length has a large influence on the results thus questioning the reliability of the extracted dynamics [4, 5, 6]. Furthermore, the lack of consensus on what drives the underlying neurological changes questions what is the appropriate model for dFC. As an alternative to windowing setting the window length to 1 and imposing smoothness in the state transitions leads to a hidden Markov model (HMM), which has been used for modeling dFC in several recent publications [7, 8, 9, 10, 11, 12].

Recently, dFC approaches have been applied in the context of schizophrenic patients and shown promise in characterizing the differences between patient and healthy controls. In fMRI-studies, the focus has been on understanding heredity of the disease [13], the disease influence on working memory [14] as well as hallucinations [15] and resting-state dFC differences between medicated patient populations and controls [16, 17, 18, 19].

In this paper, we investigate how different HMM modeling assumptions on the dynamics in resting state fMRI translate into classification accuracy using a cohort of schizophrenic patients (SZ) and healthy controls (HC). We accomplish

Corresponding author: sfvn@dtu.dk. This work was supported by the Lundbeckfonden grant no. R105-9813, the Novo Nordisk Foundation Interdisciplinary Synergy Program 2014 (BASICS) grant no. NNF14OC0011413 and by the NIH grant R01-EB-020407. We thank Qunfang Long for assistance with the group ICA.

this using the Bayesian hidden Markov model framework [7, 11, 20] with different emission models and investigate their ability to discriminate between SZ and HC. The different emission models, that each encode different assumptions on dynamics, will be compared using classification accuracy on held-out data. The purpose of this paper, then, is to use classification performance as a tool to evaluate the utility of different modeling assumptions about dynamic functional connectivity. Thus we pose the following research questions to be answered: 1) How do different assumptions on dynamic functional connectivity models influence classification performance? 2) To what extent does modeling dynamic (as opposed to static) functional connectivity influence classification performance?

2. METHODS

We use the variational Bayes hidden Markov model (VB-HMM) from [7,20] (and the accompanying MATLAB implementation[1]). The VB-HMM with K states has the generative model for the observations $\mathbf{x}_t \in \mathbb{R}^p$ for $t = 1...T$,

$$\boldsymbol{\pi}_0 \sim \mathrm{Dir}(\boldsymbol{\kappa}_0) \tag{1}$$

$$\boldsymbol{\pi}^{(k)} \sim \mathrm{Dir}(\boldsymbol{\kappa}^{(k)}), \tag{2}$$

$$z_t | z_{t-1} \sim \mathrm{Multinomial}(\boldsymbol{\pi}^{(z_{t-1})}), \tag{3}$$

$$\boldsymbol{\Sigma}^{(k)^{-1}} \sim \mathcal{W}(\boldsymbol{\Sigma}_0, \nu_0), \tag{4}$$

$$\boldsymbol{\mu}^{(k)} \sim \mathcal{N}(\boldsymbol{\mu}_0, \lambda^{-1}\boldsymbol{\Sigma}^{(k)}), \tag{5}$$

$$\mathbf{x}_t \sim \mathcal{N}(\boldsymbol{\mu}^{(z_t)}, \boldsymbol{\Sigma}^{(z_t)}), \tag{6}$$

in which $\boldsymbol{\pi}_0$ is the initial state distribution vector of length K, $\mathrm{Dir}(\cdot)$ is the Dirichlet distribution, $\boldsymbol{\kappa}_0$ is the prior vector for the initial distribution, $\boldsymbol{\pi}^{(k)}$ is a row of the transition matrix, $\boldsymbol{\kappa}^{(k)}$ is the associated prior to that row, z_t is the integer valued state taking possible values from $1...K$ at time point t, $\boldsymbol{\Sigma}^{(k)^{-1}}$ is the precision matrix from the k'th state assumed to be Wishart distributed ($\mathcal{W}$) with priors $\boldsymbol{\Sigma_0}$ and ν_0 whereas $\boldsymbol{\mu}_0$ is the prior on the mean of each state with associated scaling parameter λ.

To create a classifier from an HMM-model we use a density-based approach. For that we need the predictive likelihood on held out subjects. We use a VB-approximation [11, 21] to the predictive likelihood by calculating the free-energy on the test set keeping the transition matrix and state specific parameters fixed from training, and neglecting the terms in the free-energy that have not changed from training. This corresponds to for given training data $\mathbf{X}$ and test data $\mathbf{X}^*$ to the following bound multiplying by $\frac{Q_{X^*}(\mathbf{z}^*)}{Q_{X^*}(\mathbf{z}^*)} = 1$ and using Jensen's inequality,

$$\begin{aligned}
&\ln p(\mathbf{X}^*|\mathbf{X}) \approx \ln \int\int\int\int [p(\mathbf{X}^*\mathbf{z}^*|\boldsymbol{\pi}_0, \boldsymbol{\pi}, \boldsymbol{\theta}_{obs}) \\
&Q_X(\boldsymbol{\pi}_0)Q_X(\boldsymbol{\pi})Q_X(\boldsymbol{\theta}_{obs})]\, d\boldsymbol{\pi}_0 d\boldsymbol{\pi} d\boldsymbol{\theta}_{obs} d\mathbf{z}^* \\
&\geq \langle \ln p(\mathbf{X}^*, \mathbf{z}^*|\boldsymbol{\pi}_0, \boldsymbol{\pi}, \boldsymbol{\theta}_{obs})\rangle_{Q_X(\boldsymbol{\pi}_0)Q_X(\boldsymbol{\pi})Q_X(\boldsymbol{\theta}_{obs})Q_{X^*}(\mathbf{z}^*)} \\
&- \langle \ln Q_{X^*}(\mathbf{z}^*)\rangle_{Q_{X^*}(\mathbf{z}^*)},
\end{aligned} \tag{7}$$

in which $\boldsymbol{\theta}_{obs}$ is all the parameters in the emission model, $Q_X(\cdot)$ is the fitted variational distribution to the training set and $Q_{X^*}(\cdot)$ is the corresponding distribution for the test set, whereas $\mathbf{z}^*$ is the state sequence of the test set.

For a given training and test split we end up with two models each only trained to their respective group, $\mathcal{M}_{SZ}$ and $\mathcal{M}_{HC}$. Now we can evaluate for a new data set, X^*, what model/group was most likely to generate the data by Bayes rule, $p(\mathcal{M}_{SZ}|X^*) = \frac{p(X^*|\mathcal{M}_{SZ})p(\mathcal{M}_{SZ})}{\sum_{c=\{HC,SZ\}} p(X^*|\mathcal{M}_c)p(\mathcal{M}_c)}$, in which $p(X^*|\mathcal{M}_{SZ})$ is the predictive likelihood on test set X^* by model $\mathcal{M}_{SZ}$ and $p(\mathcal{M}_{SZ})$ is our prior of observing that model. We set this to the empirical proportions in the training data, i.e. $p(\mathcal{M}_{SZ}) = \frac{\#\mathrm{SZ}}{\#\mathrm{SZ}+\#\mathrm{HC}}$.

It is unclear what characterizes differences between SZ and HC. We therefore consider the six different emission parameterizations given in Table 1 each based on different characterizations of dFC. The diffences could be driven by changes in interaction between ICA components accounted for by having the full covariance $\boldsymbol{\Sigma}^{(k)}$ ("Mean+Cov" and "Zero-Mean"), or potentially only by within component differences not taking interactions into account ("Diag-Cov" and "Diag-Cov Zero-Mean"), or solely changes in mean activity with stationary (co-)variance ("Stationary Cov" and "Stationary Diag-Cov"). By varying the model order we further quantify if differences are best characterized by static differences between groups ($K = 1$) or relies on the dynamic characterizations ($K > 1$). We thus use the classification accuracy to quantify which parameterization best discriminates between SZ and HC.

3. RESULTS

In the following we will present the results from a synthetic study and a resting-state fMRI data set containing schizophrenic patients and healthy controls. In all of the analyses we set the priors in the HMM models to their defaults as explained in [7].

Equivalence of Different Emission Models: There are many different ways of parameterizing the underlying brain dynamics. In the six emission models we have chosen there are some equivalences in the representations which we have to take into consideration when interpreting the results. To illustrate this we have generated two data sets (mimicking two groups for classification) from two "Stationary Diag-Cov" models in Figure 1 (left panel). Both models have two states

[1]Code was downloaded from the following Github repository: `https://github.com/OHBA-analysis/HMM-MAR` in July 2016

Name [Free Parameters] (Description)	Parameterization
Mean+Cov $[K(p+p(p+1)/2)]$ (Bias and component interaction)	$\boldsymbol{\Sigma}^{(k)^{-1}} \sim \mathcal{W}(\boldsymbol{\Sigma}_0, \nu_0), \quad k = 1...K$ $\boldsymbol{\mu}^{(k)} \sim \mathcal{N}(\boldsymbol{\mu}_0, \lambda^{-1}\boldsymbol{\Sigma}^{(k)}) \quad k = 1...K$
Zero-Mean $[Kp(p+1)/2]$ (No bias but only component interaction)	$\boldsymbol{\Sigma}^{(k)^{-1}} \sim \mathcal{W}(\boldsymbol{\Sigma}_0, \nu_0), \quad k = 1...K$ $\boldsymbol{\mu}^{(k)} = \mathbf{0} \quad k = 1...K$
Diag-Cov $[2Kp]$ (Bias and within component modulation)	$\sigma_i^{(k)^{-1}} \sim \mathcal{G}(a_0, b_0), \quad i = 1..p, \quad k = 1...K$ $\boldsymbol{\Sigma}^{(k)^{-1}} = \text{diag}\left(\left[\sigma_1^{(k)^{-1}}, \sigma_2^{(k)^{-1}}, ..., \sigma_p^{(k)^{-1}}\right]\right), \quad k = 1...K$ $\boldsymbol{\mu}^{(k)} \sim \mathcal{N}(\boldsymbol{\mu}_0, \lambda^{-1}\boldsymbol{\Sigma}^{(k)}) \quad k = 1...K$
Diag-Cov Zero-Mean $[Kp]$ (No bias but only within component modulation)	$\sigma_i^{(k)} \sim \mathcal{G}(a_0, b_0), \quad i = 1...p, \quad k = 1...K$ $\boldsymbol{\Sigma}^{(k)^{-1}} = \text{diag}\left(\left[\sigma_1^{(k)^{-1}}, \sigma_2^{(k)^{-1}}, ..., \sigma_p^{(k)^{-1}}\right]\right), \quad k = 1..K$ $\boldsymbol{\mu}^{(k)} = \mathbf{0} \quad k = 1..K$
Stationary Cov $[p(p+1)/2 + Kp]$ (Bias with stationary component interaction)	$\boldsymbol{\Sigma^{-1}} \sim \mathcal{W}(\boldsymbol{\Sigma}_0, \nu_0)$ $\boldsymbol{\mu}^{(k)} \sim \mathcal{N}(\boldsymbol{\mu}_0, \lambda^{-1}\boldsymbol{\Sigma}) \quad k = 1..K$
Stationary Diag-Cov $[p + Kp]$ (Bias with stationary within component modulation)	$\sigma_i \sim \mathcal{G}(a_0, b_0), \quad i = 1..p$ $\boldsymbol{\Sigma^{-1}} = \text{diag}([\sigma_1^{-1}, \sigma_2^{-1}, ..., \sigma_p^{-1}])$ $\boldsymbol{\mu}^{(k)} \sim \mathcal{N}(\boldsymbol{\mu}_0, \lambda^{-1}\boldsymbol{\Sigma}) \quad k = 1..K$

Table 1: Overview of the six different HMM emission model parameterizations tested. The model is written for the emission space $\mathbb{R}^p$, i.e. we observe time series from $i = 1...p$ regions or independent components, and we model that with K states. The diag-operator used above takes a p-dimensional vector as input and produces a $p \times p$ matrix with the input vector in the diagonal and zeros elsewhere. Furthermore, $\mathcal{G}(a, b)$ denotes the gamma distribution.

($K = 2$) but the states differ in their mean values over the two groups making the classification task possible. However, in the bottom of Figure 1 we show the static covariance matrix for each group, i.e., equivalent to fitting the "Zero-Mean" model with one state. We notice that the classification task is still feasible since the two covariance matrices are very different even though we have a model mismatch in terms of which model generated the data.

To investigate this more systematically, we generated a synthetic dataset containing two groups with 100 subjects in each; the individual subjects data were generated with $T = 150$ timepoints (matching the data used in the subsequent analysis) in $p = 3$ dimensions. We used the state-means from the synthetic data illustrated in Figure 1 (left panel), and otherwise identical parameters across the two groups (i.e. $\boldsymbol{\pi}$, $\boldsymbol{\pi}_0$, and diagonal covariance). The data were demeaned and set to unit variance as done in the GIFT-toolbox (cf. section below). The classification accuracy obtained from 10-fold stratified cross-validation can be seen in Figure 1 (right panel). We see that all the emission-models can achieve perfect classification accuracy, except the "Diag-Cov Zero-Mean" model that is unable to account for the dynamic difference in mean activation present across the two groups.

Schizophrenia Classification We ran our analysis on a cohort consisting of 192 subjects' resting-state fMRI data (COBRE) [22]. Of those, 101 subjects were diagnosed as schizophrenic or schizoaffective (SZ) and 91 subjects were healthy controls (HC). We ran a gICA using the GIFT toolbox [23] with the ERBM algorithm [24] and 85 components. We restarted the algorithm 25 times and chose the best run using the minimum spanning tree (MST) criterion [25]. Afterwards we calculated the fractional amplitude of low-frequency fluctuation (fALFF) [26] of each component and removed all components with a fALFF lower than 3, yielding 48 components. Finally, we visually inspected the spatial maps and removed four additional components that had spatial overlap known noise sources (e.g. ventricles), such that we ended up with 44 ICs. Note that GIFT by default standardizes the time-series to have zero mean and unit variance which will become important when we compare the different model parameterizations. We estimate the accuracy of the classifiers by stratified 10-fold cross-validation. Each HMM-model was initialized 5 times, and the model with the best free-energy was chosen for the subsequent classification step. In Figure 1 we report the mean accuracy over folds, and the standard error, i.e. the standard deviation on the mean. We also report the performance of the baseline classifier, that assigns all data points in the test set to the largest class from the training set.

From the performance curves of the different HMM models we observe that all models except the "Diag-Cov Zero-Mean" have similar classification accuracies (the errorbars overlap). There seems to be a low influence on how many states we choose; there are intervals, i.e. from 4-6 states, where the more complex models "Mean+Cov" and "Zero-Mean" pull ahead in average classification accuracy, however

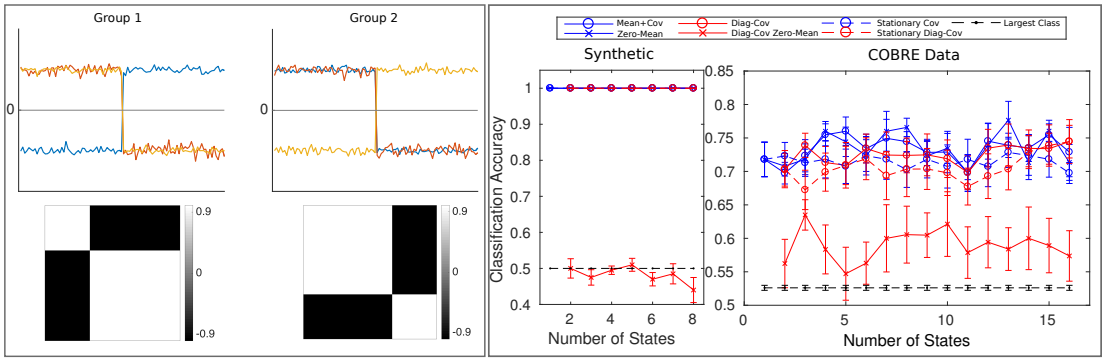


Fig. 1: *Left panel*: Synthetic toy-data showing the equivalence of different emission models. We generated two datasets in three dimensions each from a "Shared Diag-Cov" emission model (see top). Each model contains two states with a change point in the middle of the sequence. Below, we visualize the empirical correlation matrix for each data set. *Right panel*: Classification accuracy as a function of the number of states used in the different HMM-models, where the errorbars indicate the standard error over folds. Accuracy was estimated based on stratified 10-fold cross-validation. Note that the diagonal covariance models ("Diag-Cov", "Diag-Cov Zero-Mean" and "Shared Diag-Cov") start in $K = 2$ due to time-series standardization employed by the GIFT-toolbox, which makes the one state model unable to discriminate.

the errorbars still overlap with some of the simpler models.

4. DISCUSSION

In this work, we evaluated different assumptions on dFC within the HMM framework, by their ability to discriminate between schizophrenic patients (SZ) and health controls (HC) based on a short resting state fMRI scan.

Answer to research question 1: The performance gap between the full-parameterized model (with both mean and covariance for each state) and the more constrained models was fairly low. Only the "Diag-Cov Zero-Mean" model assuming that only the variance of the components varies over time, gave a noticeable drop in classification accuracy when compared to the other models. As simple models accounting for dynamic changes in the mean performed on par with models accounting for interactions between components this could indicate that the ICA we have employed as a "preprocessing" step has sufficiently demixed the problem. Thus the discriminative signal is mainly characterized by within component differences, and not in their coupling. From Figure 1 it seems that different model parameterizations can carry the same discriminative information. For example, if a certain state is characterized by one region having above mean activation and another region having below mean activation, this can be modeled in several ways in the HMM as illustrated in the synthetic data. The most natural way would be to do this using an emission model with a mean, however, a zero-mean model with full covariance could also model this by a large negative covariance between the two regions in question. Since we do not see a large discriminative effect in the more complex emission models compared to their constrained counterparts this could make the case that the differences between SZ and HC is adequately captured by non-stationary mean IC activation.

Answer to research question 2: We saw that the performance of the different models was not highly influenced by the number of states chosen in the model. This could again be an effect of the representation that we have chosen, i.e. the ICA. If all the "dynamics" are captured by the ICs and we are in a sufficiently demixed space then there is no need to subsequently fit a temporally dynamic model like the HMM beyond how these ICs are stationary coupled (i.e., the "Zero-Mean" emission model for $K = 1$).

We stress that there can be a distinction between the model that is best for *classification* and the model that best *characterizes* the data. The conclusion we make about the dynamic models here are based on their ability to discriminate between two populations, and even though we conclude that a simple emission model can bring us a long way, this does not mean that full-covariance models should be ruled out. However, care should be taken when estimating many parameters (such as full covariance matrices in the complex emission models) when data is limited. The difference between characterization and classification has to be investigated further, along with the relationship between different subspace representations, such as PCA, ICA as well as atlas parcellations into functional units, and how these representations influence the estimated dynamic functional connectivity. We argue that the presently considered predictive classification accuracy is an important complementary tool to tools quantifying models ability to characterize data [11].

5. REFERENCES

[1] R. M. Hutchison et al., "Dynamic functional connectivity: promise, issues, and interpretations," *Neuroimage*, vol. 80, pp. 360–378, Oct. 2013.

[2] V. D. Calhoun and T. Adali, "Time-Varying brain connectivity in fMRI data: Whole-brain data-driven approaches for capturing and characterizing dynamic states," *IEEE Signal Process. Mag.*, vol. 33, no. 3, pp. 52–66, May 2016.

[3] E. A. Allen et al., "Tracking whole-brain connectivity dynamics in the resting state," *Cereb. Cortex*, vol. 24, no. 3, pp. 663–676, Mar. 2014.

[4] T. O. Laumann et al., "On the stability of BOLD fMRI correlations," *Cereb. Cortex*, Sept. 2016.

[5] S. Shakil et al., "Evaluation of sliding window correlation performance for characterizing dynamic functional connectivity and brain states," *Neuroimage*, vol. 133, pp. 111–128, Mar. 2016.

[6] R. Hindriks et al., "Can sliding-window correlations reveal dynamic functional connectivity in resting-state fMRI?," *Neuroimage*, vol. 127, pp. 242–256, Feb. 2016.

[7] D. Vidaurre et al., "Spectrally resolved fast transient brain states in electrophysiological data," *Neuroimage*, vol. 126, pp. 81–95, Feb. 2016.

[8] A. P. Baker et al., "Fast transient networks in spontaneous human brain activity," *Elife*, 2014.

[9] S. Ryali et al., "Temporal dynamics and developmental maturation of salience, default and Central-Executive network interactions revealed by variational bayes hidden markov modeling," *PLoS Comput. Biol.*, vol. 12, no. 12, pp. e1005138, Dec. 2016.

[10] S. F. V. Nielsen et al., "Nonparametric modeling of dynamic functional connectivity in fMRI data," in *NIPS 2015 Workshop on Machine Learning and Interpretation in Neuroimaging*, I. Rish et al., Eds. Jan. 2016, arxiv.org.

[11] S. F. V. Nielsen et al., "Predictive assessment of models for dynamic functional connectivity," *Neuroimage*, vol. 171, pp. 116–134, Dec. 2017.

[12] J. Ou et al., "Characterizing and differentiating brain state dynamics via hidden markov models," *Brain Topogr.*, vol. 28, no. 5, pp. 666–679, Sept. 2015.

[13] J. Su et al., "Heredity characteristics of schizophrenia shown by dynamic functional connectivity analysis of resting-state functional MRI scans of unaffected siblings," *Neuroreport*, vol. 27, no. 11, pp. 843–848, Aug. 2016.

[14] C. M. Cassidy et al., "Dynamic connectivity between brain networks supports working memory: Relationships to dopamine release and schizophrenia," *J. Neurosci.*, vol. 36, no. 15, pp. 4377–4388, Apr. 2016.

[15] S. Lefebvre et al., "Network dynamics during the different stages of hallucinations in schizophrenia," *Hum. Brain Mapp.*, vol. 37, no. 7, pp. 2571–2586, July 2016.

[16] H. Shen et al., "Internetwork dynamic connectivity effectively differentiates schizophrenic patients from healthy controls," *Neuroreport*, vol. 25, no. 17, pp. 1344–1349, Dec. 2014.

[17] X. Wang et al., "Aberrant intra-salience network dynamic functional connectivity impairs large-scale network interactions in schizophrenia," *Neuropsychologia*, vol. 93, no. Pt A, pp. 262–270, Dec. 2016.

[18] Y. Du et al., "Dynamic functional connectivity impairments in early schizophrenia and clinical high-risk for psychosis," *Neuroimage*, Oct. 2017.

[19] E. Damaraju et al., "Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia," *Neuroimage Clin*, vol. 5, pp. 298–308, July 2014.

[20] D. Vidaurre et al., "Discovering dynamic brain networks from big data in rest and task," *Neuroimage*, June 2017.

[21] M. J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, Ph.D. thesis, 2003.

[22] S. Ma et al., "Dynamic changes of spatial functional network connectivity in healthy individuals and schizophrenia patients using independent vector analysis," *Neuroimage*, vol. 90, pp. 196–206, Apr. 2014.

[23] V. D. Calhoun et al., "A method for making group inferences from functional MRI data using independent component analysis," *Hum. Brain Mapp.*, vol. 14, no. 3, pp. 140–151, Nov. 2001.

[24] X. L. Li and T. Adali, "Blind spatiotemporal separation of second and/or higher-order correlated sources by entropy rate minimization," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. Mar. 2010, pp. 1934–1937, ieeexplore.ieee.org.

[25] W. Du et al., "The role of diversity in complex ICA algorithms for fMRI analysis," *J. Neurosci. Methods*, vol. 264, pp. 129–135, May 2016.

[26] Q.-H. Zou et al., "An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: fractional ALFF," *J. Neurosci. Methods*, vol. 172, no. 1, pp. 137–141, July 2008.

A.5 Testing Group Differences in State Transition Structure of Dynamic Functional Connectivity Models

Nielsen, Søren F V, Diego Vidaurre, Kristoffer H Madsen, Mikkel N Schmidt, and Morten Mørup (2018). "Testing group differences in state transition structure of dynamic functional connectivity models". In: *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*. Singapore.

Testing group differences in state transition structure of dynamic functional connectivity models

Søren F.V. Nielsen*, Diego Vidaurre‡, Kristoffer H. Madsen*,†, Mikkel N. Schmidt* and Morten Mørup*
* DTU Compute, Technical University of Denmark, Denmark
†Danish Research Centre for Magnetic Resonance, Copenhagen University Hospital Hvidovre, Denmark
‡ OHBA, Welcome Centre for Integrative Neuroimaging, Department of Psychiatry, University of Oxford, UK
Corresponding email: sfvn@dtu.dk

***Abstract*—Understanding the origins of intrinsic time-varying functional connectivity remains a challenge in the neuroimaging community. However, some associations between dynamic functional connectivity (dFC) and behavioral traits have been observed along with gender differences. We propose a permutation testing framework to investigate dynamic differences between groups of subjects. In particular, we investigate differences in fractional occupancy, state persistency and the full transition probability matrix. We demonstrate our framework on resting state functional magnetic resonance imaging data from 820 healthy young adults from the Human Connectome Project considering two prominent dFC models, namely sliding-window k-means and the Gaussian hidden Markov model. The variables showing consistent significant dynamic differences were limited to gender and the degree of motion in the scanner. We observe for the data considered that a large sample size (here 500 subjects) is needed to to draw reliable conclusions about the significance of those variables. Our results point to dynamic features providing limited information with regard to behavioral traits despite a relatively large sample size.**

I. Introduction

Neuroimaging has over the last decade moved from localizing brain function, mainly using statistical parametric mapping (SPM), and into characterizing *functional connectivity* (FC), i.e. the statistical dependencies between segregated brain regions. Especially in resting-state (rs) functional magnetic resonance imaging (fMRI) a lot of research papers have investigated how we can explain FC-differences in healthy populations [1], [2] and how we can use FC as biomarkers for neuropsychiatric diseases [3].

More recently, the implicit/explicit assumption of temporal stationary FC in rs-fMRI has been questioned and investigated [4], [5], [6], [7], which has fueled the modeling of so-called dynamic FC (dFC) *states*. dFC states are a discrete set of FC patterns that reoccur in time, both within subjects [8], [9] and across a population [5], [10]. The two most prominent methods for modeling dFC states is the sliding-window k-means (SWKM) [5] and the Gaussian hidden Markov model (HMM) [11], [10].

Recently, the temporal characteristics of dFC and their relation to cognitive measures has been investigated. Ma et al. [12] investigated the use of SWKM on rs-fMRI data from a cohort of patients diagnosed with schizophrenia and healthy controls. They found qualitative differences between the two groups in the so-called *transition matrix* estimated post-hoc from dFC states. The transition matrix quantifies for all time steps the probability of switching between any two states. Ma et al. [12] did not apply statistical testing of the transition differences.

Vidaurre et al. [10] trained a Gaussian hidden Markov model (HMM) on rs-fMRI from healthy young adults in the Human Connectome Project (HCP). The HMM is a probabilistic generative model of the data that assumes a latent discrete state space which is 1st order Markovian. Here the transition matrix is estimated directly from the HMM fitting procedure. They found that the states extracted had a hierarchical structure in terms of time each subject spent in each state, denoted fractional occupancy (FO). Two meta-states from the top of the hierarchy were then extracted and the difference in FO between the two meta-states, called the meta-state profile, associated with behavioral data significantly better than random (obtained through permutation testing). Furthermore, a comparison of the transition matrix in subgroups defined by their meta-state FO was carried out and showed qualitative differences. Vidaurre et al. [10] did however not carry out a quantitative analysis of the relationship between transition features and behavioral data.

In this paper, we propose a permutation framework for testing for group differences in the transition dynamics of dFC. We apply the framework to the rs-fMRI data from the Human Connectome Project [13] considering both SWKM and HMM. We do this by first training a dFC model (HMM and SWKM) on the entire population (820 subjects), and subsequently we estimate the transition matrix in two subgroups based on behavioural data (gender, motion, personality traits, etc). To characterise the difference between the two transition matrices we use the total variation (TV) distance of probability measures. We further contrast the performance to simple properties of the dFC models given by fractional

*Søren F.V. Nielsen, Mikkel N. Schmidt and Morten Mørup were supported by Lundbeckfonden (fellowship grant R105-9813 to Morten Mørup). Kristoffer H. Madsen was supported by a Novo Nordisk Foundation Interdisciplinary Synergy Grant (NNF14OC0011413). Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

978-1-5386-4291-7/18/$31.00 ©2018 IEEE

occupancy (FO) and a measure of self-transitions denoted global state persistency (GSP). In particular we investigate; 1) What behavioral variables significantly influence transition dynamics and how influenced is transition dynamics of head motion? 2) What aspects of the transition dynamics are important for characterizing these differences? 3) In this population, how many subjects are necessary to reliably detect group differences?

II. Methods

A. Dynamic functional connectivity state transitions

Dynamic functional connectivity (dFC) models in general describe the changes in temporal correlation between two brain regions (i and j), c_{ij}. Thus at each time point, $t = 1 \dots T$, we have a snapshot of the FC between all pairs of regions, $\mathbf{C}^{(t)}$. A dFC *state* model further assumes that the $\mathbf{C}^{(t)}$'s can be clustered into K states, which yields a *state sequence*, $\mathbf{z}$, i.e. the assignment of each time point to one of the K states, $z_t \in \{1, 2, \dots, K\}$.

Assume that we have obtained a state sequence from a dFC state model (e.g. HMM or SWKM), then the K-by-K transition matrix, $\mathbf{P}$, can be written as,

$$\mathbf{P}_{k,k'} = \frac{\sum_{t=2}^{T} \delta(z_t = k', z_{t-1} = k)}{\sum_{t=2}^{T} \delta(z_{t-1} = k)}. \tag{1}$$

The element $\mathbf{P}_{k,k'}$ thus quantifies the probability of transitioning from state k to state k'. Furthermore, each row is a probability distribution meaning that it sums to one, $\sum_{k'} \mathbf{P}_{k,k'} = 1$.

We also quantify the overall persistency of all the states, which we will denote the *global state persistency* (GSP), by taking the mean of the diagonal of the transition matrix, i.e.

$$\mathrm{GSP} = \frac{\sum_j \mathbf{P}_{j,j}}{K}. \tag{2}$$

Finally, we consider fractional occupancy FO_k which is a commonly used statistic to characterize clustering solutions [10], which can be calculated as,

$$\mathrm{FO}_k = \frac{\sum_t \delta(z_t = k)}{T}, \tag{3}$$

i.e. this quantifies how much of the total time is spent in the state k. Notice that this also sums to 1 over states and thereby forms a probability distribution, however it disregards the temporal structure of the state sequence.

B. Permutation testing using group information

To assess statistical differences between the dFC transition features of two groups we use approximate nonparametric permutation testing [14]. We investigate dFC transitions at a population level where we have data from S subjects, where each subject's state sequence can be denoted $\mathbf{z}^{(s)}$ for $s = 1 \dots S$. This state sequence is obtained by a population-level analysis, i.e. all subjects (regardless of grouping) have been concatenated into one long sequence. Given the grouping information, $\mathbf{g} \in [1, 2]^S$, we want to post-hoc estimate the difference in transition patterns between the groups. Each group's transition matrix is estimated on the collection of state sequences, i.e. group 1 has the transition matrix $\mathbf{P}^{(1)}$ estimated from $\mathbf{Z}^{(\mathbf{1})} = \{\mathbf{z}^{(s)} \; \forall \; s \; : \; \mathbf{g}_s = 1\}$. Another approach would be to train the dFC model with a transition matrix for each group, however, this approach is computationally expensive as we would need to retrain the model for each permutation.

As a distance measure between the transition probability matrices we use the total variation measure (TV) summed over the rows of $\mathbf{P}$. The TV between two probability distributions corresponds to the largest difference in probability which the two distributions assigns to the same event [15]. Another way to measure "closeness" is the Kullback-Leibler (KL) divergence, which is related to TV through Pinsker's inequality. However, this has the disadvantage that it is not symmetric and degenerates when an element has zero probability mass (due to a logarithm). We investigated using a regularized version of KL instead of TV with no difference in the conclusions of this paper. The TV distance can be written as,

$$\mathrm{TV}(\mathbf{P}^{(1)}, \mathbf{P}^{(2)}) = \sum_{k=1}^{K} \frac{1}{2} \sum_{j=1}^{K} \left| \mathbf{P}^{(1)}_{k,j} - \mathbf{P}^{(2)}_{k,j} \right|, \tag{4}$$

where $|\cdot|$ is the absolute value.

This same measure can be applied to the fractional occupancies (FO) for each group. For the GSP measure we take the absolute value of the difference between the two groups' GSP. For the permutation testing we permute the group labels and reestimate the transition matrices (and FO) for the permuted groups and calculate the distance between them. We thereby obtain a null-distribution of the considered measure between the groups by repeating the procedure for a large number of permutations as defined by the smallest p-value obtainable [16]. We used 10^5 permutations for our main analysis, which lets us obtain a minimum p-value of 10^{-5}, and used Bonferroni-correction with correction-factor equal to the number of behavioral variables ($m = 10$).

III. Results

We investigate the above permutation testing framework on resting state (rs) functional magnetic resonance imaging (fMRI) data from the Human Connectome Project (HCP) 820 subject release [13]. The data has been parcellated into 50 components using a group independent component analysis (ICA) publicly available through the HCP website[1]. Data were temporally concatenated and standardised such that each IC time-course within a subject had zero mean and unit variance. Afterwards, we ran the variational Bayes hidden Markov model (HMM) using the `HMM-MAR` MATLAB-toolbox[2] with $K = 12$ states and the stochastic inference engine [11]. All states had individual mean and full covariance in order to be comparable to the analysis carried out by [10]. We in addition ran the sliding-window k-means (SWKM) with the same number of states as the HMM ($K = 12$) using

[1] https://db.humanconnectome.org/

[2] https://github.com/OHBA-analysis/HMM-MAR

a window of length $W = [60, 100, 150]$ convolved with a Gaussian ($\sigma = 3TR$) [5] sliding the window one TR at a time. We did not use shorter window lengths because this necessitates regularization of the correlation matrix, such as the sparse-inverse regularization approach from [5], which was too computationally demanding. For the k-means inference we used the `litekmeans` implementation [17]. We investigated grouping the subjects into two groups according to 10 behavioral variables; gender, the five factor traits [18], two self reported measures of stress, fluid intelligence and a measure of head motion estimated from the realignment procedure (session average). All of the continuous variables were thresholded to match the proportions in the gender variable. We discarded four subjects that had missing values among the behavioral data we chose to investigate.

The results of the permutation testing for the fractional occupancy (FO), global state persistency (GSP) and transition probability matrix (TPM) (as calculated by (4)) can be seen in Figure 1. In general gender and motion (unsurprisingly) yield significant differences in almost all of the measures and models with few interesting exceptions. Looking at the results for the HMM we note that in the FO the largest difference is observed for the motion variable whereas in the transition matrix this is true for the gender variable. And looking at the GSP it is only the gender variable that overall shows significant differences. For the SWKM (three rightmost columns of the figure) we observe very similar results for the short window lengths ($W = [60, 100]$), however when we increase the window length to $W = 150$ the GSP no longer shows significant results for any of the behavioral variables. Furthermore, FO and TPM differences for Gender and Motion move closer to the tail of the null-distribution as compared to the shorter window lengths.

IV. Discussion

The neural origins of dFC in resting state fMRI is still not very well understood. To what extent it is best explained by cognitive differences in the subjects, ongoing cognitive processing, anatomical differences or noise confounds remains an open question. In this paper we have presented a framework for investigating dFC transition differences in groups of subjects using permutation testing. We applied this to healthy adults' resting state fMRI data from the Human Connectome Project.

Overall, we found no statistical evidence to support dFC differences in groups defined by higher-order cognitive and psychological traits (such as the five factor model) in neither of the dFC models considered (HMM and SWKM). However, we acknowledge that the thresholding we have applied to the continuous variables reduces the resolution, such that detection of transition differences is no longer possible. Gender showed significant differences in (almost) all of our analyses; however, this was expected since sex differences in anatomy are quite large, which could lead to systematic differences in the BOLD signal [19]. Recently, a machine learning model based on neuroanatomical features was trained on 967 subjects was trained to predict their gender, and achieved a 86% (cross-validated) classification accuracy [20]. Our analysis revealed that grouping subjects by how much they moved inside the scanner also gave significant differences in dFC features. This is also fairly unsurprising as motion has been put forward as a strong bias in discovering behavior and static FC relationships [21]. Furthermore, in the domain of dFC head motion has been attributed the strongest source of dFC variance by Laumann et al [22]. We investigated three dFC temporal features derived from the state sequence. FO and TPM differences were significant for gender and motion, whereas for GSP gender was the only significant variable across both HMM and SWKM. This indicates that head motion influences transitions to new states and overall time spent in particular states more compared to state persistency.

Our empirical investigation into the power of the permutation testing framework shows that we need quite a lot of subjects (> 500) to get reliable significant differences in the transition probability matrices (cf. bottom of Figure1). However, the absolute differences between the elements of the TPMs between males and females were very low (on the order of 10^{-3}). This shows that the effect is very small but reliable enough to be detectable in the large sample size. Future work will include using prediction of continuous behavioral variables on held-out subjects to investigate and disentangle FC and dFC features.

References

[1] S. M. Smith et al., "A positive-negative mode of population covariation links brain connectivity, demographics and behavior," *Nat. Neurosci.*, vol. 18, no. 11, pp. 1565–1567, Nov. 2015.

[2] S. Smith, "Linking cognition to brain connectivity," *Nat. Neurosci.*, vol. 19, no. 1, pp. 7–9, Jan. 2016.

[3] T. Yamada et al., "Resting-State functional Connectivity-Based biomarkers and functional MRI-Based neurofeedback for psychiatric disorders: A challenge for developing theranostic biomarkers," *Int. J. Neuropsychopharmacol.*, vol. 20, no. 10, pp. 769–781, Oct. 2017.

[4] C. Chang and G. H. Glover, "Time-frequency dynamics of resting-state brain connectivity measured with fMRI," *Neuroimage*, vol. 50, no. 1, pp. 81–98, Mar. 2010.

[5] E. A. Allen et al., "Tracking whole-brain connectivity dynamics in the resting state," *Cereb. Cortex*, vol. 24, no. 3, pp. 663–676, Mar. 2014.

[6] R. M. Hutchison et al., "Dynamic functional connectivity: promise, issues, and interpretations," *Neuroimage*, vol. 80, pp. 360–378, Oct. 2013.

[7] V. D. Calhoun et al., "The chronnectome: time-varying connectivity networks as the next frontier in fMRI data discovery," *Neuron*, vol. 84, no. 2, pp. 262–274, Oct. 2014.

[8] J. Gonzalez-Castillo et al., "Tracking ongoing cognition in individuals using brief, whole-brain functional connectivity patterns," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 28, pp. 8762–8767, July 2015.

[9] S. F. V. Nielsen et al., "Predictive assessment of models for dynamic functional connectivity," *Neuroimage*, vol. 171, pp. 116–134, Dec. 2017.

[10] D. Vidaurre et al., "Brain network dynamics are hierarchically organized in time," *Proc. Natl. Acad. Sci. U. S. A.*, Oct. 2017.

[11] D. Vidaurre et al., "Discovering dynamic brain networks from big data in rest and task," *Neuroimage*, June 2017.

[12] S. Ma et al., "Dynamic changes of spatial functional network connectivity in healthy individuals and schizophrenia patients using independent vector analysis," *Neuroimage*, vol. 90, pp. 196–206, Apr. 2014.

[13] S. M. Smith et al., "Resting-state fMRI in the human connectome project," *Neuroimage*, vol. 80, pp. 144–168, Oct. 2013.

[14] T. E. Nichols and A. P. Holmes, "Nonparametric permutation tests for functional neuroimaging: a primer with examples," *Hum. Brain Mapp.*, vol. 15, no. 1, pp. 1–25, Jan. 2002.

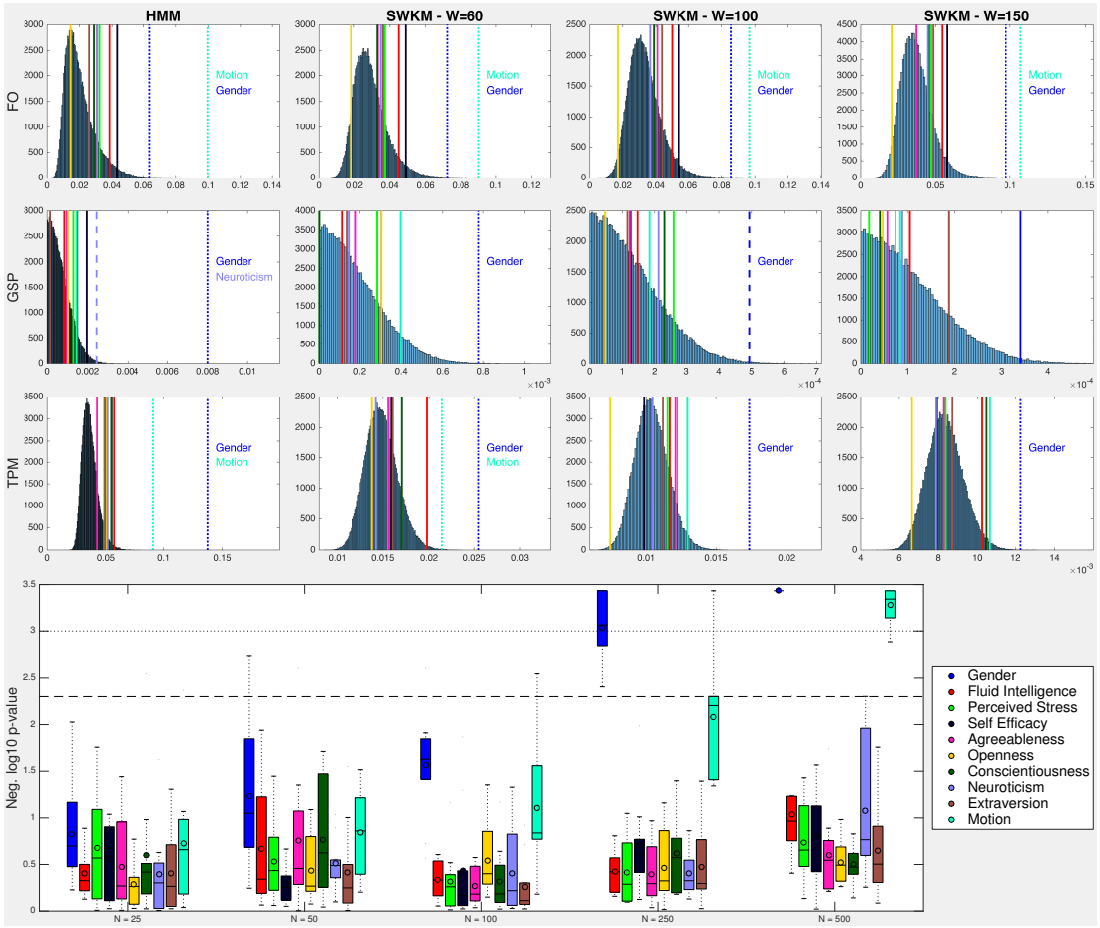


Fig. 1: *Top figure:* Permutation results for the SWKM and HMM using 10^5 permutations in the null-distribution. The SWKM and HMM were each retrained 10 times and we chose the best model according to the minimum cost function value. Behavioral variables were split to match the proportions in the *Gender* variable, such that hypothesis testing could be carried out using the same null-distribution for all behaviorals. For each behavioral it has been indicated by the linestyle if significant differences were detected (Bonferroni corrected with $\alpha = 0.05$ for dashed lines and $\alpha = 0.01$ for dotted lines). The significant ($\alpha = 0.05$ level) variable names are indicated in each subplot. *Bottom figure:* We analyse the influence of the number of subjects ($N = [25, 50, 100, 250, 500]$) on the permutation framework. Using the best HMM solution (out of 10 restarts), trained on the entire HCP820 data, we calculated TV on the TPMs for different groups (described above) and we report the estimated p-value using 10^4 permutations. The boxplot above is the negative log p-value over 10 random subsets (H0: grouping yields same TV). The dashed black line indicates significance level $\alpha = 0.05$ and the dotted line significance level $\alpha = 0.01$ (both Bonferroni-corrected).

[15] D. A. Levin and Y. Peres, *Markov Chains and Mixing Times: Second Edition*, American Mathematical Soc., Oct. 2017.

[16] A. M. Winkler et al., "Permutation inference for the general linear model," *Neuroimage*, vol. 92, pp. 381–397, May 2014.

[17] D. Cai et al., "Document clustering using locality preserving indexing," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 12, pp. 1624–1637, Dec. 2005.

[18] J. M. Digman, "Personality structure: Emergence of the Five-Factor model," *Annu. Rev. Psychol.*, vol. 41, no. 1, pp. 417–440, Jan. 1990.

[19] B. B. Biswal et al., "Toward discovery science of human brain function," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 10, pp. 4734–4739, Mar. 2010.

[20] F. Sepehrband et al., "Neuroanatomical morphometric characterization of sex differences in youth using statistical learning," *Neuroimage*, vol. 172, pp. 217–227, Feb. 2018.

[21] J. S. Siegel et al., "Data quality influences observed links between functional connectivity and behavior," *Cereb. Cortex*, vol. 27, no. 9, pp. 4492–4502, Sept. 2017.

[22] T. O. Laumann et al., "On the stability of BOLD fMRI correlations," *Cereb. Cortex*, Sept. 2016.

Appendix B

Technical Appendix

B.1 Predictive Likelihood in HMM using MCMC

The predictive likelihood of unseen data is a quantity of interest when choosing between different hidden Markov models (HMM) (cf. section 2.3). The predictive likelihood can be written as,

$$p(\boldsymbol{X}^*|\boldsymbol{X}) = \int\int\int \sum_{\mathbf{z}^*} p(\boldsymbol{X}^*, \mathbf{z}^*|\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\pi}_0, \boldsymbol{X}) p(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\pi}_0|\boldsymbol{X}) d\boldsymbol{\theta} d\boldsymbol{\pi} d\boldsymbol{\pi}_0, \quad \text{(B.1)}$$

in which $\boldsymbol{X}$ is the training set, $\boldsymbol{X}^*$ is the test set, $\mathbf{z}^*$ denotes the test state sequence, $\boldsymbol{\theta}$ is the collection of all parameters relevant to the emission model, $\boldsymbol{\pi}$ is the transition matrix and $\boldsymbol{\pi}_0$ is the vector containing the initial state probabilities. Equation (B.1) can be approximated using the S samples obtained from the MCMC training procedure such that,

$$p(\boldsymbol{X}^*|\boldsymbol{X}) \approx \frac{1}{S} \sum_s \sum_{\mathbf{z}^*} p(\boldsymbol{X}^*, \mathbf{z}^*|\boldsymbol{\theta}^{(s)}, \boldsymbol{\pi}^{(s)}, \boldsymbol{\pi}_0^{(s)}, \boldsymbol{X}), \quad \text{(B.2)}$$

in which $\boldsymbol{\theta}^{(s)}, \boldsymbol{\pi}^{(s)}, \boldsymbol{\pi}_0^{(s)}$ denote parameter samples from the MCMC procedure.

The summation over all possible state sequences can be handled efficiently using modification of the classical Viterbi-algorithm (Viterbi, 1967). To derive this, a

simple dataset is considered only containing two data points, x_1 and x_2, and a posterior sample of $\boldsymbol{\theta}$, $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_0$ obtained using a two-state HMM. The number of all possible state sequences in this case is four and we can thus write the summation out as,

$$\begin{aligned} p([x_1, x_2]|\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\pi}_0) &= \sum_{\mathbf{z}^*} p([x_1, x_2], \mathbf{z}^*|\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\pi}_0) \\ &= p(x_1|z_1 = 1)p(z_1 = 1)p(x_2|z_2 = 1)p(z_2 = 1|z_1 = 1) \\ &+ p(x_1|z_1 = 1)p(z_1 = 1)p(x_2|z_2 = 2)p(z_2 = 2|z_1 = 1) \\ &+ p(x_1|z_1 = 2)p(z_1 = 2)p(x_2|z_2 = 1)p(z_2 = 1|z_1 = 2) \\ &+ p(x_1|z_1 = 2)p(z_1 = 2)p(x_2|z_2 = 2)p(z_2 = 2|z_1 = 2), \end{aligned}$$

in which the explicit conditional parameter dependencies have been left out for notational ease (i.e. $p(z_1 = 1|\boldsymbol{\pi}_0) = p(z_1 = 1)$). The likelihood terms for x_2 can be collected and rearranging yields,

$$\begin{aligned} p([x_1, x_2]|\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\pi}_0) &= p(x_2|z_2 = 1)\,[p(x_1|z_1 = 1)p(z_1 = 1)p(z_2 = 1|z_1 = 1) \\ &+ p(x_1|z_1 = 2)p(z_1 = 2)p(z_2 = 1|z_1 = 2)] \\ &+ p(x_2|z_2 = 2)\,[p(x_1|z_1 = 1)p(z_1 = 1)p(z_2 = 2|z_1 = 1) \\ &+ p(x_1|z_1 = 2)p(z_1 = 2)p(z_2 = 2|z_1 = 2)] \\ &= p(x_2|z_2 = 1)V_{2,1} + p(x_2|z_2 = 2)V_{2,2}, \\ V_{2,k} &= \sum_j p(x_1|z_1 = j)p(z_1 = j)p(z_2 = k|z_1 = j), \quad k = 1, 2 \end{aligned}$$

The summation has a particular structure that can be exploited. At each timestep the likelihood from the previous timestep is accumulated and weighted by the proper transition probabilities (which are given by $\boldsymbol{\pi}$). The values in V can thus be estimated recursively. For a general dataset of length T and an HMM with K states we can obtain desired summation as,

$$\sum_{\mathbf{z}^*} p(\boldsymbol{X}^*, \mathbf{z}^*|\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\pi}_0) = \sum_{k=1}^{K} V_{T,k},$$

in which

$$V_{1,k} = p(x_1|z_1 = k)p(z_1 = k), \quad k = 1..K \tag{B.3}$$

$$V_{t,k} = p(x_t|z_t = k) \sum_{j=1}^{K} p(z_t = k|z_{t-1} = j)V_{t-1,j}, \quad t = 2..T, k = 1..K \tag{B.4}$$

In practice, the predictive log-likelihood is often calculated due to numerical stability properties. In that case the "log-sum-exp"-trick is used for calculating the summations in the above formulas.

Bibliography

Akaike, H (1974). "A new look at the statistical model identification". In: *IEEE Trans. Automat. Contr.* 19.6, pp. 716–723. ISSN: 0018-9286. DOI: 10.1109/TAC.1974.1100705.

Allen, Elena A, Eswar Damaraju, Sergey M Plis, Erik B Erhardt, Tom Eichele, and Vince D Calhoun (2014). "Tracking whole-brain connectivity dynamics in the resting state". In: *Cereb. Cortex* 24.3, pp. 663–676. ISSN: 1047-3211, 1460-2199. DOI: 10.1093/cercor/bhs352.

Andersen, Kasper Winther, Kristoffer H Madsen, Hartwig Roman Siebner, Mikkel N Schmidt, Morten Mørup, and Lars Kai Hansen (2014). "Non-parametric Bayesian graph models reveal community structure in resting state fMRI". In: *Neuroimage* 100, pp. 301–315. ISSN: 1053-8119, 1095-9572. DOI: 10.1016/j.neuroimage.2014.05.083.

Baum, Leonard E (1972). "An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes". In: *Inequalities III: Proceedings of the Third Symposium on Inequalities*. Ed. by Oved Shisha. University of California, Los Angeles: Academic Press, pp. 1–8.

Beal, Matthew J (2003). "Variational Algorithms for Approximate Bayesian Inference". PhD thesis.

Beal, Matthew J, Zoubin Ghahramani, and Carl E Rasmussen (2002). "The Infinite Hidden Markov Model". In: *Advances in Neural Information Processing Systems 14*. Ed. by T. G. Dietterich and S. Becker and Z. Ghahramani. MIT Press, pp. 577–584.

Bergmeir, Christoph, Rob J Hyndman, and Bonsoo Koo (2018). "A note on the validity of cross-validation for evaluating autoregressive time series pre-

diction". In: *Comput. Stat. Data Anal.* 120, pp. 70–83. ISSN: 0167-9473. DOI: 10.1016/j.csda.2017.11.003.

Bishop, Christopher M (1999). "Variational Principal Components". In: *Proceedings of the 1999 the 9th International Conference on 'Artificial Neural Networks (ICANN99)*. IEEE, pp. 509–514.

– (2006). *Pattern recognition and machine learning*. Springer.

Biswal, B, F Z Yetkin, V M Haughton, and J S Hyde (1995). "Functional connectivity in the motor cortex of resting human brain using echo-planar MRI". In: *Magn. Reson. Med.* 34.4, pp. 537–541. ISSN: 0740-3194.

Blei, David M, Alp Kucukelbir, and Jon D McAuliffe (2016). "Variational Inference: A Review for Statisticians". In: arXiv: 1601.00670 [stat.CO].

Bzdok, Danilo and B T Thomas Yeo (2017). "Inference in the age of big data: Future perspectives on neuroscience". en. In: *Neuroimage*. ISSN: 1053-8119, 1095-9572. DOI: 10.1016/j.neuroimage.2017.04.061.

Calhoun, Vince D, Tulay Adali, G D Pearlson, and J J Pekar (2001). "A method for making group inferences from functional MRI data using independent component analysis". en. In: *Hum. Brain Mapp.* 14.3, pp. 140–151. ISSN: 1065-9471.

Calhoun, Vince D, Robyn Miller, Godfrey Pearlson, and Tulay Adalı (2014). "The chronnectome: time-varying connectivity networks as the next frontier in fMRI data discovery". In: *Neuron* 84.2, pp. 262–274. ISSN: 0896-6273, 1097-4199. DOI: 10.1016/j.neuron.2014.10.015.

Chang, Catie and Gary H Glover (2010). "Time-frequency dynamics of resting-state brain connectivity measured with fMRI". In: *Neuroimage* 50.1, pp. 81–98. ISSN: 1053-8119, 1095-9572. DOI: 10.1016/j.neuroimage.2009.12.011.

Cherian, Anoop, Vassilios Morellas, and Nikolaos Papanikolopoulos (2016). "Bayesian Nonparametric Clustering for Positive Definite Matrices". en. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 38.5, pp. 862–874. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2015.2456903.

Choe, Ann S, Mary Beth Nebel, Anita D Barber, Jessica R Cohen, Yuting Xu, James J Pekar, Brian Caffo, and Martin A Lindquist (2017). "Comparing test-retest reliability of dynamic functional connectivity methods". en. In: *Neuroimage* 158, pp. 155–175. ISSN: 1053-8119, 1095-9572. DOI: 10.1016/j.neuroimage.2017.07.005.

De Castro, Yohann, É, and Claire Lacour (2016). "Minimax Adaptive Estimation of Nonparametric Hidden Markov Models". In: *J. Mach. Learn. Res.* 17.111, pp. 1–43. ISSN: 1532-4435.

Du, Yuhui, Susanna L Fryer, Zening Fu, Dongdong Lin, Jing Sui, Jiayu Chen, Eswar Damaraju, Eva Mennigen, Barbara Stuart, Daniel H Mathalon, and Vince D Calhoun (2017). "Dynamic functional connectivity impairments in early schizophrenia and clinical high-risk for psychosis". en. In: *Neuroimage*. ISSN: 1053-8119, 1095-9572. DOI: 10.1016/j.neuroimage.2017.10.022.

Fox, Emily B (2009). "Bayesian nonparametric learning of complex dynamical phenomena". PhD thesis. Massachusetts Institute of Technology.

Fox, Emily B, Erik B Sudderth, Michael I Jordan, and Alan S Willsky (2008). "An HDP-HMM for systems with state persistence". In: *Proceedings of the 25th international conference on Machine learning*. ACM, pp. 312–319. ISBN: 9781605582054. DOI: 10.1145/1390156.1390196.

Fox, Michael D and Marcus E Raichle (2007). "Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging". en. In: *Nat. Rev. Neurosci.* 8.9, pp. 700–711. ISSN: 1471-003X. DOI: 10.1038/nrn2201.

Friston, Karl J (2011). "Functional and effective connectivity: a review". In: *Brain Connect.* 1.1, pp. 13–36. ISSN: 2158-0014, 2158-0022. DOI: 10.1089/brain.2011.0008.

Friston, Karl J, Lee Harrison, and Will Penny (2003). "Dynamic causal modelling". In: *Neuroimage* 19.4, pp. 1273–1302. ISSN: 1053-8119. DOI: 10.1016/S1053-8119(03)00202-7.

Friston, Karl J, Cathy J Price, Paul Fletcher, Caroline Moore, Richard S Frackowiak, and Raymond J Dolan (1996). "The trouble with cognitive subtraction". en. In: *Neuroimage* 4.2, pp. 97–104. ISSN: 1053-8119. DOI: 10.1006/nimg.1996.0033.

Geisser, Seymour and William F Eddy (1979). "A Predictive Approach to Model Selection". In: *J. Am. Stat. Assoc.* 74.365, pp. 153–160. ISSN: 0162-1459. DOI: 10.1080/01621459.1979.10481632.

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin (2014). *Bayesian data analysis*. Vol. 2. CRC press Boca Raton, FL.

Gonzalez-Castillo, Javier, Colin W Hoy, Daniel A Handwerker, Meghan E Robinson, Laura C Buchanan, Ziad S Saad, and Peter A Bandettini (2015). "Tracking ongoing cognition in individuals using brief, whole-brain functional connectivity patterns". In: *Proc. Natl. Acad. Sci. U. S. A.* 112.28, pp. 8762–8767. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1501242112.

Good, Irving John (1952). "Rational decisions". In: *J. R. Stat. Soc. Series B Stat. Methodol.* Pp. 107–114. ISSN: 1369-7412.

Greene, Abigail S, Siyuan Gao, Dustin Scheinost, and R Todd Constable (2018). "Task-induced brain state manipulation improves prediction of individual traits". en. In: *Nat. Commun.* 9.1, p. 2807. ISSN: 2041-1723. DOI: 10.1038/s41467-018-04920-3.

Handwerker, Daniel A, Vinai Roopchansingh, Javier Gonzalez-Castillo, and Peter A Bandettini (2012). "Periodic changes in fMRI connectivity". en. In: *Neuroimage* 63.3, pp. 1712–1719. ISSN: 1053-8119, 1095-9572. DOI: 10.1016/j.neuroimage.2012.06.078.

Hidot, Sullivan and Christophe Saint-Jean (2010). "An Expectation–Maximization algorithm for the Wishart mixture model: Application to movement clustering". In: *Pattern Recognit. Lett.* 31.14, pp. 2318–2324. ISSN: 0167-8655. DOI: 10.1016/j.patrec.2010.07.002.

Hoffman, Matthew D, David M Blei, Chong Wang, and John Paisley (2013). "Stochastic variational inference". In: *J. Mach. Learn. Res.* 14.1, pp. 1303–1347. ISSN: 1532-4435.

Hutchison, R Matthew, Thilo Womelsdorf, Elena A Allen, Peter A Bandettini, Vince D Calhoun, Maurizio Corbetta, Stefania Della Penna, Jeff H Duyn, Gary H Glover, Javier Gonzalez-Castillo, Daniel A Handwerker, Shella Keilholz, Vesa Kiviniemi, David A Leopold, Francesco de Pasquale, Olaf Sporns, Martin Walter, and Catie Chang (2013). "Dynamic functional connectivity: promise, issues, and interpretations". In: *Neuroimage* 80, pp. 360–378. ISSN: 1053-8119, 1095-9572. DOI: 10.1016/j.neuroimage.2013.05.079.

Jain, Sonia and Radford M Neal (2004). "A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model". In: *J. Comput. Graph. Stat.* 13.1, pp. 158–182. ISSN: 1061-8600. DOI: 10.1198/1061860043001.

Kandasamy, Kirthevasan, Maruan Al-Shedivat, and Eric P Xing (2016). "Learning HMMs with Nonparametric Emissions via Spectral Decompositions of Continuous Matrices". In: *Advances in Neural Information Processing Systems 29*. Ed. by D D Lee, M Sugiyama, U V Luxburg, I Guyon, and R Garnett. Curran Associates, Inc., pp. 2865–2873.

Korzen, Josefine, Kristoffer H Madsen, and Morten Mørup (2014). *Quantifying Temporal States in rs-fMRI Data using Bayesian Nonparametrics*. Poster presentation at Human Brain Mapping 2014. Hamburg, Germany.

Kucyi, Aaron (2017). "Just a thought: How mind-wandering is represented in dynamic brain connectivity". en. In: *Neuroimage*. ISSN: 1053-8119, 1095-9572. DOI: 10.1016/j.neuroimage.2017.07.001.

Kwong, K K, J W Belliveau, D A Chesler, I E Goldberg, R M Weisskoff, B P Poncelet, D N Kennedy, B E Hoppel, M S Cohen, and R Turner (1992). "Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation". In: *Proc. Natl. Acad. Sci. U. S. A.* 89.12, pp. 5675–5679. ISSN: 0027-8424.

Laumann, Timothy O, Abraham Z Snyder, Anish Mitra, Evan M Gordon, Caterina Gratton, Babatunde Adeyemo, Adrian W Gilmore, Steven M Nelson, Jeff J Berg, Deanna J Greene, John E McCarthy, Enzo Tagliazucchi, Helmut Laufs, Bradley L Schlaggar, Nico U F Dosenbach, and Steven E Petersen (2016). "On the Stability of BOLD fMRI Correlations". en. In: *Cereb. Cortex*. ISSN: 1047-3211, 1460-2199. DOI: 10.1093/cercor/bhw265.

Leonardi, Nora and Dimitri Van De Ville (2015). "On spurious and real fluctuations of dynamic functional connectivity during rest". en. In: *Neuroimage* 104, pp. 430–436. ISSN: 1053-8119, 1095-9572. DOI: 10.1016/j.neuroimage.2014.09.007.

Liégeois, Raphaël, Timothy O Laumann, Abraham Z Snyder, Juan Zhou, and B T Thomas Yeo (2017). "Interpreting temporal fluctuations in resting-state functional connectivity MRI". en. In: *Neuroimage*. ISSN: 1053-8119, 1095-9572. DOI: 10.1016/j.neuroimage.2017.09.012.

Ma, Sai, Vince D Calhoun, Ronald Phlypo, and Tülay Adalı (2014). "Dynamic changes of spatial functional network connectivity in healthy individuals and schizophrenia patients using independent vector analysis". In: *Neuroimage* 90, pp. 196–206. ISSN: 1053-8119, 1095-9572. DOI: 10.1016/j.neuroimage.2013.12.063.

Miller, Robyn L, Tulay Adali, Yuri Levin-Schwartz, and Vince D Calhoun (2017). "Resting-State fMRI Dynamics and Null Models: Perspectives, Sampling Variability, and Simulations". en.

Neal, Radford M (1993). *Probabilistic inference using Markov chain Monte Carlo methods*. Tech. rep. University of Toronto.

Nielsen, Søren F V (2015). "Modelling Dynamic Functional Brain Connectivity". MA thesis. Technical University of Denmark.

Nielsen, Søren F V, Yuri Levin-Schwartz, Diego Vidaurre, Tulay Adali, Vince D Calhoun, Kristoffer H Madsen, Lars Kai Hansen, and Morten Mørup (2018). "Evaluating models of dynamic functional connectivity using predictive classification accuracy". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Calgary, AB, Canada.

Nielsen, Søren F V, Kristoffer H Madsen, Rasmus Røge, Mikkel N Schmidt, and Morten Mørup (2016). "Nonparametric Modeling of Dynamic Functional Connectivity in fMRI Data". In: *NIPS 2015 Workshop on Machine Learning and Interpretation in Neuroimaging*. Montreal, Canada: arxiv.org.

Nielsen, Søren F V, Kristoffer H Madsen, Mikkel N Schmidt, and Morten Mørup (2017). "Modeling dynamic functional connectivity using a wishart mixture model". In: *2017 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*. Toronto, Canada: IEEE, pp. 1–4. DOI: 10.1109/PRNI.2017.7981505.

Nielsen, Søren F V, Mikkel N Schmidt, Kristoffer H Madsen, and Morten Mørup (2018). "Predictive assessment of models for dynamic functional connectivity". en. In: *Neuroimage* 171, pp. 116–134. ISSN: 1053-8119, 1095-9572. DOI: 10.1016/j.neuroimage.2017.12.084.

Nielsen, Søren F V, Diego Vidaurre, Kristoffer H Madsen, Mikkel N Schmidt, and Morten Mørup (2018). "Testing group differences in state transition structure of dynamic functional connectivity models". In: *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*. Singapore.

Ogawa, S, D W Tank, R Menon, J M Ellermann, S G Kim, H Merkle, and K Ugurbil (1992). "Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging". In: *Proc. Natl. Acad. Sci. U. S. A.* 89.13, pp. 5951–5955. ISSN: 0027-8424.

Ou, Jinli, Li Xie, Changfeng Jin, Xiang Li, Dajiang Zhu, Rongxin Jiang, Yaowu Chen, Jing Zhang, Lingjiang Li, and Tianming Liu (2015). "Characterizing and Differentiating Brain State Dynamics via Hidden Markov Models". en. In: *Brain Topogr.* 28.5, pp. 666–679. ISSN: 0896-0267, 1573-6792. DOI: 10.1007/s10548-014-0406-2.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay (2011). "Scikit-learn: Machine Learning in Python". In: *J. Mach. Learn. Res.* 12.Oct, pp. 2825–2830. ISSN: 1532-4435, 1533-7928.

Piironen, Juho and Aki Vehtari (2017). "Comparison of Bayesian predictive methods for model selection". In: *Stat. Comput.* 27.3, pp. 711–735. ISSN: 0960-3174, 1573-1375. DOI: 10.1007/s11222-016-9649-y.

Poldrack, Russell A, Timothy O Laumann, Oluwasanmi Koyejo, Brenda Gregory, Ashleigh Hover, Mei-Yen Chen, Krzysztof J Gorgolewski, Jeffrey Luci, Sung Jun Joo, Ryan L Boyd, Scott Hunicke-Smith, Zack Booth Simpson, Thomas Caven, Vanessa Sochat, James M Shine, Evan Gordon, Abraham Z Snyder, Babatunde Adeyemo, Steven E Petersen, David C Glahn, D Reese Mckay, Joanne E Curran, Harald H H Göring, Melanie A Carless, John Blangero, Robert Dougherty, Alexander Leemans, Daniel A Handwerker, Laurie Frick, Edward M Marcotte, and Jeanette A Mumford (2015). "Long-term neural and physiological phenotyping of a single human". In: *Nat. Commun.* 6, p. 8885. ISSN: 2041-1723. DOI: 10.1038/ncomms9885.

Pourahmadi, Mohsen and Xiao Wang (2015). "Distribution of random correlation matrices: Hyperspherical parameterization of the Cholesky factor". In: *Stat. Probab. Lett.* 106, pp. 5–12. ISSN: 0167-7152.

Qin, Shaozheng, Christina B Young, Kaustubh Supekar, Lucina Q Uddin, and Vinod Menon (2012). "Immature integration and segregation of emotion-related brain circuitry in young children". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 109.20, pp. 7941–7946. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1120408109.

Rabiner, L R (1989). "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Proc. IEEE* 77.2, pp. 257–286. ISSN: 0018-9219. DOI: 10.1109/5.18626.

Raichle, Marcus E, A M MacLeod, A Z Snyder, W J Powers, D A Gusnard, and G L Shulman (2001). "A default mode of brain function". In: *Proc. Natl. Acad. Sci. U. S. A.* 98.2, pp. 676–682. ISSN: 0027-8424. DOI: 10.1073/pnas.98.2.676.

Rasmussen, Peter M, Lars K Hansen, Kristoffer H Madsen, Nathan W Churchill, and Stephen C Strother (2012). "Model sparsity and brain pattern interpretation of classification models in neuroimaging". In: *Pattern Recognit.* 45.6, pp. 2085–2100. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2011.09.011.

Rezek, Iead and Stephen Roberts (2005). "Ensemble Hidden Markov Models with Extended Observation Densities for Biosignal Analysis". en. In: *Probabilistic Modeling in Bioinformatics and Medical Informatics*. Ed. by Dirk Husmeier, Richard Dybowski, and Stephen Roberts. Advanced Information and Knowledge Processing. Springer London, pp. 419–450. DOI: 10.1007/1-84628-119-9_14.

Ryali, Srikanth, Kaustubh Supekar, Tianwen Chen, John Kochalka, Weidong Cai, Jonathan Nicholas, Aarthi Padmanabhan, and Vinod Menon (2016). "Temporal Dynamics and Developmental Maturation of Salience, Default and Central-Executive Network Interactions Revealed by Variational Bayes Hidden Markov Modeling". en. In: *PLoS Comput. Biol.* 12.12, e1005138. ISSN: 1553-734X, 1553-7358. DOI: 10.1371/journal.pcbi.1005138.

Sakoğlu, Unal, Godfrey D Pearlson, Kent A Kiehl, Y Michelle Wang, Andrew M Michael, and Vince D Calhoun (2010). "A method for evaluating dynamic functional network connectivity and task-modulation: application to schizophrenia". en. In: *MAGMA* 23.5-6, pp. 351–366. ISSN: 0968-5243, 1352-8661. DOI: 10.1007/s10334-010-0197-8.

Schwarz, Gideon (1978). "Estimating the Dimension of a Model". en. In: *Ann. Stat.* 6.2, pp. 461–464. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1176344136.

Scott, Steven L, Alexander W Blocker, Fernando V Bonassi, Hugh A Chipman, Edward I George, and Robert E McCulloch (2016). "Bayes and big data: the consensus Monte Carlo algorithm". In: *International Journal of Management Science and Engineering Management* 11.2, pp. 78–88. ISSN: 1750-9653. DOI: 10.1080/17509653.2016.1142191.

Shakil, Sadia, Chin-Hui Lee, and Shella Dawn Keilholz (2016). "Evaluation of sliding window correlation performance for characterizing dynamic functional connectivity and brain states". In: *Neuroimage* 133, pp. 111–128. ISSN: 1053-8119, 1095-9572. DOI: 10.1016/j.neuroimage.2016.02.074.

Smith, Stephen M, Christian F Beckmann, Jesper Andersson, Edward J Auerbach, Janine Bijsterbosch, Gwenaëlle Douaud, Eugene Duff, David A Feinberg, Ludovica Griffanti, Michael P Harms, Michael Kelly, Timothy Laumann, Karla L Miller, Steen Moeller, Steve Petersen, Jonathan Power, Gholamreza Salimi-Khorshidi, Abraham Z Snyder, An T Vu, Mark W Woolrich, Junqian Xu, Essa Yacoub, Kamil Uğurbil, David C Van Essen, Matthew F Glasser, and WU-Minn HCP Consortium (2013). "Resting-state fMRI in the Human Connectome Project". In: *Neuroimage* 80, pp. 144–168. ISSN: 1053-8119, 1095-9572. DOI: 10.1016/j.neuroimage.2013.05.039.

Smith, Stephen M, Peter T Fox, Karla L Miller, David C Glahn, P Mickle Fox, Clare E Mackay, Nicola Filippini, Kate E Watkins, Roberto Toro, Angela R Laird, and Christian F Beckmann (2009). "Correspondence of the brain's functional architecture during activation and rest". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 106.31, pp. 13040–13045. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0905267106.

Taghia, Jalil, Weidong Cai, Srikanth Ryali, John Kochalka, Jonathan Nicholas, Tianwen Chen, and Vinod Menon (2018). "Uncovering hidden brain state dynamics that regulate performance and decision-making during cognition". In: *Nat. Commun.* 9.1, p. 2505. ISSN: 2041-1723. DOI: 10.1038/s41467-018-04723-6.

Vidaurre, Diego, Romesh Abeysuriya, Robert Becker, Andrew J Quinn, Fidel Alfaro-Almagro, Stephen M Smith, and Mark W Woolrich (2017). "Discovering dynamic brain networks from big data in rest and task". en. In: *Neuroimage*. ISSN: 1053-8119, 1095-9572. DOI: 10.1016/j.neuroimage.2017.06.077.

Vidaurre, Diego, Andrew J Quinn, Adam P Baker, David Dupret, Alvaro Tejero-Cantero, and Mark W Woolrich (2016). "Spectrally resolved fast transient brain states in electrophysiological data". en. In: *Neuroimage* 126, pp. 81–95. ISSN: 1053-8119, 1095-9572. DOI: 10.1016/j.neuroimage.2015.11.047.

Vidaurre, Diego, Stephen M Smith, and Mark W Woolrich (2017). "Brain network dynamics are hierarchically organized in time". en. In: *Proc. Natl. Acad. Sci. U. S. A.* ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1705120114.

Viterbi, A (1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". In: *IEEE Trans. Inf. Theory* 13.2, pp. 260–269. ISSN: 0018-9448. DOI: 10.1109/TIT.1967.1054010.

Wakeman, Daniel G and Richard N Henson (2015). "A multi-subject, multi-modal human neuroimaging dataset". en. In: *Sci Data* 2, p. 150001. ISSN: 2052-4463. DOI: 10.1038/sdata.2015.1.

Watanabe, Sumio (2010). "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory". In: *J. Mach. Learn. Res.* 11.Dec, pp. 3571–3594. ISSN: 1532-4435, 1533-7928.

Yaesoubi, M, R L Miller, T Adali, and V D Calhoun (2016). "Time-varying frequency modes of resting fMRI brain networks reveal significant gender differences". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ieeexplore.ieee.org, pp. 6310–6314. DOI: 10.1109/ICASSP.2016.7472891.

Zalesky, Andrew and Michael Breakspear (2015). "Towards a statistical test for functional connectivity dynamics". en. In: *Neuroimage* 114, pp. 466–470. ISSN: 1053-8119, 1095-9572. DOI: 10.1016/j.neuroimage.2015.03.047.

Zou, Qi-Hong, Chao-Zhe Zhu, Yihong Yang, Xi-Nian Zuo, Xiang-Yu Long, Qing-Jiu Cao, Yu-Feng Wang, and Yu-Feng Zang (2008). "An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: fractional ALFF". en. In: *J. Neurosci. Methods* 172.1, pp. 137–141. ISSN: 0165-0270. DOI: 10.1016/j.jneumeth.2008.04.012.